Prediction of Drug–Target Interactions From Multi-Molecular Network Based on Deep Walk Embedding Model

Zhan-Heng Chen1,2, Zhu-Hong You1,2+, Zhen-Hao Guo1,2, Hai-Cheng Yi1,2, Gong-Xu Luo1,2 and Yan-Bin Wang3

1 The Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences, Urumqi, China, 2 University of Chinese Academy of Sciences, Beijing, China, 3 School of Cyber Science and Technology, Zhejiang University, Hangzhou, China

Predicting drug–target interactions (DTIs) is crucial in innovative drug discovery, drug repositioning and other fields. However, there are many shortcomings for predicting DTIs using traditional biological experimental methods, such as the high-cost, time-consumption, low efficiency, and so on, which make these methods difficult to widely apply. As a supplement, the in silico method can provide helpful information for predictions of DTIs in a timely manner. In this work, a deep walk embedding method is developed for predicting DTIs from a multi-molecular network. More specifically, a multi-molecular network, also called molecular associations network, is constructed by integrating the associations among drug, protein, disease, lncRNA, and miRNA. Then, each node can be represented as a behavior feature vector by using a deep walk embedding method. Finally, we compared behavior features with traditional attribute features on an integrated dataset by using various classifiers. The experimental results revealed that the behavior feature could be performed better on different classifiers, especially on the random forest classifier. It is also demonstrated that the use of behavior information is very helpful for addressing the problem of sequences containing both self-interacting and non-interacting pairs of proteins. This work is not only extremely suitable for predicting DTIs, but also provides a new perspective for the prediction of other biomolecules’ associations.

Keywords: drug–target interactions, molecular association network, attribute feature, behavior feature, random forest

INTRODUCTION

Prediction of drug–target interactions (DTIs) is one of the most important steps in the genomic drug discovery pipeline and drug repurposing (Knowles and Gromo, 2003; Yildirim et al., 2007), the purpose is to discover putative new drugs and new uses of existing drugs. To our knowledge, the effects of many useful protein targets on drugs are modulated by interacting with ligands, including enzymes, ion channels, G protein-coupled receptors and nuclear receptors (Yamanishi et al., 2010). The development of rapid sequencing technology and the implementation of the human genome project, which has produced massive amounts of biological data, has given birth to a new
and more large and diverse data were collected at multiple levels of the system biology. Barabasi and Oltvai (2004) proposed network biology to understand the cell's functional organization. Network biology refers to studying the biosystem network using mathematical methods and graph theory, and the network topology model. The studies have shown that cellular networks obey the general rules of network science, and it is helpful for understanding the interactions between molecules inside a living cell. Afterward, inspired by deep learning and word embedding technology in natural language processing (NLP), vector representation of nodes in automatic learning networks has become a research hotspot (Goldberg and Levy, 2014; Pennington et al., 2014; Peters et al., 2018; Devlin et al., 2018; Yang et al., 2019). This work has been gradually applied to the field of bioinformatics.

To summarize, Guo et al. (2019) for the first time proposed a MAN by integrating the associations among miRNA, lncRNA, protein, drug, and disease, where any kind of potential associations can be predicted. In this paper, we constructed a biomolecular relationship network, which contains nine kinds of associations with five types of molecules. All the molecules in the MAN were treated as nodes and all the relationships were regarded as edges. The associations between a node and other nodes in the complex network were called the behavior of the node. This work introduced two kinds of important information: the original attribute information of node itself (e.g., sequences of proteins, molecular fingerprints of drugs) and behavior information of the biomolecules. Then, a comparative experiment was carried out with a random forest (RF) classifier. The experiment results show that the behavior of the node contains more useful information than the attribute of the node in the DTIs prediction, and better results can be obtained.

RESULTS AND DISCUSSION

In order to illustrate that the behavior features of nodes contain more useful information than the traditional attribute features of biomolecules, we compared the performances of various well-known classifiers based on these two different types of features under five-fold cross-validation in various evaluation criteria. Cross-validation is mainly used to prevent over-fitting caused by over-complicated models. It is a statistical method used to evaluate the generalization ability of training data. For the five-fold cross-validation, the original data is randomly divided into five parts, and four parts are selected as the training set each time, and the remaining one part is used as the test set. The cross-validation was repeated five times, and the average value for the accuracy of the five runs was taken as the evaluation index of the final model. In this work, the number of the five training sets is 17,770, 17,770, 17,770, 17,770, 17,776, respectively; the number of five test sets is 4444, 4444, 4444, 4444, 4448, respectively.

Performance Evaluation With Support Vector Machine on Two Different Features

In the experiment, we employed the state-of-the-art method Support Vector Machine (SVM) to assess the performance...
TABLE 1 | Performance evaluation with SVM on attribute features.

| 5-folds | Acc (%) | TPR (%) | TNR (%) | PPV (%) | MCC (%) |
|---------|---------|---------|---------|---------|---------|
| 1       | 66.16   | 66.79   | 65.53   | 65.96   | 32.32   |
| 2       | 66.22   | 66.16   | 66.29   | 66.25   | 32.45   |
| 3       | 66.49   | 67.64   | 65.35   | 66.12   | 33.00   |
| 4       | 67.06   | 67.69   | 66.43   | 66.84   | 34.12   |
| 5       | 66.74   | 67.37   | 66.11   | 66.53   | 33.49   |
| Average | 66.53 ± 0.37 | 67.13 ± 0.66 | 65.94 ± 0.48 | 66.34 ± 0.35 | 33.08 ± 0.75 |

TABLE 2 | Performance evaluation with SVM on behavior features.

| 5-folds | Acc (%) | TPR (%) | TNR (%) | PPV (%) | MCC (%) |
|---------|---------|---------|---------|---------|---------|
| 1       | 74.71   | 71.56   | 77.86   | 76.37   | 49.51   |
| 2       | 77.12   | 72.73   | 81.50   | 79.72   | 54.44   |
| 3       | 75.83   | 75.07   | 76.60   | 76.23   | 51.67   |
| 4       | 75.99   | 75.83   | 76.15   | 76.07   | 51.98   |
| 5       | 75.51   | 73.41   | 77.60   | 76.62   | 51.06   |
| Average | 75.83 ± 0.87 | 73.72 ± 1.73 | 77.94 ± 2.11 | 77.00 ± 1.53 | 51.73 ± 1.79 |

The ROC curve of the SVM classifier based on attribute feature and behavior feature with 5-fold cross-validation is shown in Figures 1, 2, respectively. It is clear that the average of AUC is 0.7028 by using attribute information, the average of AUC is 0.8188 by using behavior information based on MAN network. Hence, the behavior information of nodes play an important role in the DTIs predictions.

Performance Evaluation With Random Forest on Two Different Features

In order to illustrate that the behavior features are indeed better than the attribute features, either on a single liner classifier or on an ensemble classifier, we also implemented the RF model on our experiment. In this experiment, we set the same parameters to compare the performances of the two different features on the model, the results are shown in Tables 3, 4.
The ROC curves of the RF classifier based on attribute feature and behavior feature with five-fold cross-validation are shown in Figures 3, 4, respectively. It is obvious that the average of AUC is 0.8779 by using attribute information, the average of AUC is 0.9206 by using behavior information based on the MAN. So, the behavior information of nodes play an important role in the DTI predictions.

As mentioned above, it is apparent that the constructed MAN network can receive accurate DTI detection because more behavior information can be obtained from the complex biomolecular associations network. The presented complex network has made an indelible contribution to the prediction of DTIs. The main innovations can be summed up in the following two aspects: (1) Construction of the MAN network, which integrates five types of biomolecules and nine known relationships between them. It can provide a novel potential helpful tool for predicting new DTIs across the whole field of bioinformatics; (2) Behavior features were obtained by deep walk network embedding method, which can further optimize the performance of classifiers. This method can achieve more helpful information in the data than traditional attribute features. In a few words, experimental results revealed that our presented network is not only extremely suitable for DTI prediction, but also fit for other biomolecule associations prediction.

### MATERIALS AND METHODS

#### Datasets Construction

In this article, the heterogeneous data input to the MAN is collected from nine known relationships: DTIs, drug–disease
Drug and target provides a new perspective for predicting interactions between them. The construction of the systematic MAN network and lncRNA), and edges correspond to associations among five types of biomolecules (drug, protein, disease, miRNA, lncRNA), which were listed in Table 5. These known relationships were also based on five types of biomolecules: drug, protein, disease, lncRNA, miRNA; which were shown in Figure 5. Based on the known associations, some biomolecules are suggested to interact with each other. In the MAN, a homogeneous network was constructed by disease) at the cellular level. On the assumption that there is a multi-molecular network, also called MAN by linking databases which are mentioned above, we constructed five types of biomolecules annotated in many well-known databases which are listed in Table 6. The MAN contained topological relationships and distributions among all the molecules in the heterogeneous network. Considering the local and global connection modes, this work describes the basic context and intrinsic connection profiles for the whole nodes. Therefore, the prediction of DTIs can be determined by the connection relationships of the other nodes in the network.

Multi-Molecular Network

From the collection of nine known relationships between five types of biomolecules annotated in many well-known databases which are mentioned above, we constructed a multi-molecular network, also called MAN by linking two arbitrary association nodes. The complex MAN is shown in Figure 5. 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 the systematic MAN network provides a new perspective for predicting interactions between drug and target.

### Traditional Attribute Representation

**Drug Molecular Fingerprint**

The drug molecular data was extracted from DrugBank database. To further process these data better, we calculated the Morgan fingerprints of drug molecules with the RDKit (Landrum, 2013) tool in python. The main idea of the molecular fingerprint method is that molecular structure is encoded as many substructure fingerprints in a series of binary bits, and a kernel is then applied to a molecule to generate a bit vector or count vector. Substructure pattern matching can be done using query molecules built from SMARTS which is first determined as a predefined dictionary (Guba et al., 2015). As we all know, there is a SMARTS-based implementation of the 166 public MACCS keys (Cereto-Massagué et al., 2015). As shown in Figure 6, each fingerprint bit corresponds to a fragment of the molecule, if its corresponding known fragment appears in the given molecule, the corresponding bit in the fingerprint is set to 1; otherwise, it is set to 0. Thus, each molecule can be represented as a Boolean array. In this method, although the whole molecule was divided into a great many of fragments, it still retains all the complexity of drug molecules.

**Protein Sequence**

The total protein sequence information was collected from the STRING database. For protein sequences, 20 types of amino acids were classified into four categories by the polarity of the side chain information, which contained (Ala, Val, Leu, Ile, Met, Phe, Trp, Pro), (Gly, Ser, Thr, Cys, Asn, Gln, Tyr), (Arg, Lys, His), and (Asp, Glu). Similarly, each protein sequence was transformed into a 64-dimensional (4 × 4 × 4) feature vector by counting the frequency of every subsequence appearing in the whole protein sequence, and each dimension of the vector is the normalized frequency of the corresponding 3-mer in the sequence (Rizk et al., 2013).

**Network Embedding—DeepWalk**

In 2014, Perozzi et al. (2014) proposed DeepWalk, which can learn latent representation of vertices in a network. Analogous to word2vec, it uses the co-occurrence relationship among the whole nodes in the graph to learn the vector representation of nodes. There are two stages in the process of the deepwalk method: (1) A sequence of nodes is constructed. The locally associated training data is obtained by applying a random walk generator for sampling from each node in the homogeneous network. Then, to obtain a sequence for each node by imitating the process of text generation; (2) The Skip-Gram is used to train the sampling data, and the discrete nodes are represented as vectors in the network, and the Hierarchical Softmax is used to classify the ultra-large-scale classification.

**Generation of Sequence of Nodes**

In the MAN, a homogeneous network was constructed by five research objects (miRNA, lncRNA, drug, protein, and disease) at the cellular level. On the assumption that there is a network graph $G$ a random vertex $v_t$ is uniformly sampled as the root of the random walk. Then, a walk samples uniformly from each vertex to the adjacent nodes until it

### TABLE 5 | Nine known relationships in the molecular associations network.

| Relationship | Database | Number |
|-------------|----------|--------|
| Drug–target | DrugBank (Wishart et al., 2017) | 11107 |
| Drug–disease | CTD (Davis et al., 2018) | 18416 |
| Protein–disease | DisGeNET (Piñero et al., 2016) | 25087 |
| IncRNA–target | LncRNA2Target (Cheng et al., 2018) | 690 |
| IncRNA–disease | LncRNA2Disease (Chen et al., 2012) | 1264 |
| miRNA–target | miRTarBase (Chou et al., 2017) | 4944 |
| miRNA–disease | HMDD (Huang et al., 2018) | 16427 |
| miRNA–lncRNA | lncRNASNP2 (Miao et al., 2017) | 8374 |
| Protein–protein | STRING (Szklarczyk et al., 2016) | 19237 |
| Total | N/A | 105546 |

### TABLE 6 | 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 |
reaches the maximum length. In this way, the process of text generation is simulated to find sequence information for each node in the network, e.g., $V_{14} \rightarrow V_{11} \rightarrow V_{12} \rightarrow V_{13}$, $V_{27} \rightarrow V_{23} \rightarrow V_{24} \rightarrow V_{21} \rightarrow V_{22}$, $V_{34} \rightarrow V_{32} \rightarrow V_{36} \rightarrow V_{31} \rightarrow V_{37}$, and so on. Random walks on MAN is shown in Figure 7. Afterward, the sequence of each node will be treated as a sentence in NLP as input of word2vec, and the vector representation of nodes is obtained.

**Skip-Gram Model**

Skip-Gram is one type of the word2vec model, which was proposed by McCormick (2016). It uses nodes to conjecture context, and learns vector representation by maximizing the co-occurrence probability of words within a window, and ignores the order in which nodes appear in sentences. The representation of nodes with the same context is similar. The higher the frequency of two nodes appearing in a sequence at the same time, the higher the similarity between the two nodes. The co-occurrence probability can be transformed into the product of conditional probability according to independence assumption, which can be summarized as follow:

$$P(v_{i-c}, \ldots, v_{i+c}|v_i) = \prod_{k=i-c, k \neq i}^{i+c} P(v_k|\Phi(v_i))$$  \hspace{1cm} (1)$$

where, $v_{i-c}$ and $v_{i+c}$ are the left and right context of the word $v_i$, $c$ is the size of the window. In addition, we map each vertex $v_k$ to its current representation vector $\Phi(v_k) \in \mathbb{R}^d$.

The conditional probability of each vertex in the sequence is calculated, that is, the log value of the probability of other nodes in the sequence when the node appears, and the vector representation of the node is updated with the help of the stochastic gradient descent algorithm.
Classification Models

Classification is one of the important tasks in data mining. The so-called classification is to classify the unknown data into existing categories according to its characteristics or attributes. That is to say, using given categories and known training data to learn classification rules and classifiers, and then predicting the unknown data.

Support Vector Machines

Support Vector Machine (SVM) is a supervised machine learning algorithm, which is mainly used for binary classification problems (Suykens and Vandewalle, 1999). In this algorithm, each data was considered as one point in $n$-dimensional space ($n$ is the number of features), and each eigenvalue is a value of a specific coordinate. Then, classification is carried out by finding the hyper-planes that distinguish the two classes. In the sample space, the partition of hyper-planes can be described by the following linear equations:

$$f(x) = w^T x + a = 0$$

Assuming that it has completed the separation of samples and the labels of the two samples are $[+1, -1]$, for a classifier, $f(x) > 0$ represents the class that label is $+1$, otherwise, it is $-1$. In order to maximize the distance between the nearest two classes of samples on both sides of the plane, we need to find two hyper-planes parallel to and equal to the hyper-plane.

$$f(x) = w^T x + a = +1$$

$$f(x) = w^T x + a = -1$$

Then, to maximize the interval between these two hyper-planes $max(1/||w||)$. Thus, SVM can provide a good generalization ability for classification problems.

Random Forest

Random forest is a relatively novel machine learning model. In the 1980s, Breiman (2017) developed the classification tree, which achieved classification and regression by repeating binary data, and the amount of calculation was greatly reduced. In 2001, Breiman combined classification trees into RFs, which randomized the use of variables (columns) and data (rows) to generate many classification trees, and then summarized the results of all the classification trees (Breiman, 2001). Random forest contains many decision trees in the forest, but there is no correlation between these trees. When a new sample is input to the forest, each decision tree will judge which category the sample should belong to. And then, the sample was predicted to be of the most selected category.

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

$$Gini_i = \sum_{m=1}^{[M]} \sum_{m' \neq m} p_{nm} p_{nm'} = 1 - \sum_{m=1}^{[M]} p_{nm}^2$$

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

$$J(\Phi) = - \log P(u_j|\Phi(v_k))$$
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: 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:

\[
Acc = \frac{TP + TN}{TP + FP + TN + FN}
\]  
(7)

\[
TPR = \frac{TP}{TP + FN}
\]  
(8)

\[
TNR = \frac{TN}{FP + TN}
\]  
(9)

\[
PPV = \frac{TP}{TP + FP}
\]  
(10)

\[
MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}}
\]  
(11)

where \(TP\) is the count of true interacting pairs correctly predicted, i.e., the number of true positives. \(FP\) refers to the quantity of false positives, which is described as the number of true non-interacting pairs falsely predicted. \(TN\) means the quantity of true negatives, in other words, it represents the number of true non-interacting pairs predicted correctly. \(FN\) represents the quantity of false negatives, i.e., the true interacting pairs falsely predicted to be non-interacting pairs. According to these parameters, a Receiver Operating Characteristic (ROC) was plotted to evaluate the performance of the random projection method. Then we can calculate the AUC to assess the performance of the model.

CONCLUSION
In this study, we investigated the relationship among drug, protein, miRNA, LncRNA and disease. Then, we developed a novel method to discover the potential interaction between drug and target on a large scale. We constructed a novel scheme based on the above five molecules and nine relationships arbitrarily between two molecules, which is called the MAN network. By focusing 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 the deepwalk network embedding method). To our knowledge, this is the first report to predict DTIs from a complex heterogeneous network in an overall view at the cellular level. Experimental results demonstrated that our model has achieved good prediction results, which is a new attempt to predict DTIs. This work would have potential applications for drug discovery and repositioning.

DATA AVAILABILITY STATEMENT
The raw data required to reproduce these findings cannot be shared at this time as the data also forms part of an ongoing study. Requests to access the datasets should be directed to the corresponding author.

AUTHOR CONTRIBUTIONS
Z-HC and Z-HY conceived the algorithm, carried out analyses, prepared the data sets, carried out experiments, and wrote the manuscript. Z-HG and H-CY designed and performed the experiments. G-XL and Y-BW analyzed the experiments and checked the manuscript. All authors read and approved the final manuscript.

FUNDING
This work is supported in part by the National Natural Science Foundation of China, under Grants 61373086 and 61572506.

ACKNOWLEDGMENTS
The authors would like to thank all the guest editors and reviewers for their constructive advices.

REFERENCES
Barabasi, A. L., and Oltvai, Z. N. (2004). Network biology: understanding the cell’s functional organization. Nat. Rev. Genet. 5, 101–113. doi: 10.1038/nrg1272
Breiman, L. (2001). Random forests.
Breiman, L. (2017). Classification and Regression Trees. Abingdon: Routledge.
Cahan, P., Li, H., Morris, S. A., Lummertz da Rocha, E., Daley, G. Q., and Collins, J. J. (2014). CellNet: network biology applied to stem cell engineering. Cell 158, 903–915. doi: 10.1016/j.cell.2014.07.020
Cereto-Massagüé, A., Ojeda, M. I., Valls, C., Mulero, M., Garcia-Vallvé, S., and Pujadas, G. (2015). Molecular fingerprint similarity search in virtual screening. Methods 71, 58–63. doi: 10.1016/j.ymeth.2014.08.005
Chan, K. C., and You, Z.-H. (2016). “Large-scale prediction of drug-target interactions from deep representations,” in Proceedings of the Neural Networks (IJCNN), 2016 International Joint Conference on, Vancouver, BC: IEEE.
Chen, G., Wang, Z., Wang, D., Qiu, C., Liu, M., Chen, X., et al. (2012). LncRNADisease: a database for long-non-coding RNA-associated diseases. Nucleic Acids Res. 41, D983–D986. doi: 10.1093/nar/gks1099
Chen, X., Yan, C. C., Zhang, X., Zhang, X., Dai, F., Yin, J., et al. (2015). Drug-target interaction prediction: databases, web servers and computational models. Brief. Bioinform. 17, 696–712. doi: 10.1093/bib/bbv066
Cheng, L., Wang, P., Tian, R., Wang, S., Guo, Q., Luo, M., et al. (2018). LncRNA2Target v2.0: a comprehensive database for target genes of lncRNAs in human and mouse. Nucleic Acids Res. 47, D140–D144. doi: 10.1093/nar/gky1051
Chou, C.-H., Shrestha, S., Yang, C. D., Chang, N. W., Lin, Y. L., Liao, K. W., et al. (2017). miRtarBase update 2018: a resource for experimentally validated microRNA-target interactions. Nucleic Acids Res. 46, D296–D302. doi: 10.1093/nar/gkx1067
Pennington, J., Socher, R., and Manning, C. (2014). “Glove: global vectors for word representation,” in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Stroudsburg, PA: Association for Computational Linguistics.