Peer-reviewed version available at *Cells* **2019**, *8*, 866; doi:10.3390/cells8080866

Article

Construction and Comprehensive Analysis of a Molecular Associations Network via lncRNA-miRNA -disease-drug-protein Graph

Zhen-Hao Guo [1,2,†], Hai-Cheng Yi [1,2,†], Zhu-Hong You [1,2,†,*]

[1] The Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences, Urumqi 830011, China; guozhenhao17@mails.ucas.ac.cn (Z-H. G.); yihaicheng17@mails.ucas.ac.cn (H-C. Y.)
[2] University of Chinese Academy of Sciences, Beijing 100049, China;
* Correspondence: zhuhongyou@ms.xjb.ac.cn (Z-H. Y.);
† These authors contributed equally to this work.

Abstract: The key issue in the post-genomic era is how to systematically describe the association between small molecule transcripts or translations inside cells. With the rapid development of high-throughput "omics" technologies, the achieved ability to detect and characterize molecules with other molecule targets opens up the possibility of investigating the relationships between different molecules from a global perspective. In this article, a Molecular Associations Network(MAN) is constructed and comprehensively analyzed by integrating the associations among miRNA, lncRNA, protein, drug, and disease, in which any kind of potential associations can be predicted. More specifically, each node in MAN can be represented as a vector by combining two kinds of information including the attributes of the node itself (e.g. sequences of ncRNAs and proteins, semantics of diseases and molecular fingerprints of drugs) and the manner of the node in the complex network (associations with other nodes). Random Forest classifier is trained to classify and predict new interactions or associations between biomolecules. In the experiment, the proposed method achieves a superb performance with 0.9735 AUC in 5-fold cross-validation, which show that the proposed method can provide new insight for exploration of the molecular mechanisms of disease and valuable clues for disease treatment.

Keywords: Network biology; LINE; lncRNA; protein; miRNA; Drug; disease.

1. Introduction

There are many types of biomolecules inside living cells that form multiple associated regulatory networks as pathways or direct participants to maintain a wide variety of life activities and key functions [1-3]. For instance, protein-protein interactions play a key role in numerous life processes and maintain many of the functions of normal cells. There is also growing evidence that ncRNAs are involved in cell growth and apoptosis, leading to many diseases. Therefore, predicting the potential associations between small molecule transcripts and compounds not only helps people to understand important cell activities at the molecular level, but is also significant for prevention, diagnosis and treatment of disease, as well as genomic drug discovery [4, 5]. In fact, it is unrealistic to verify the existence of association between such large-scale nodes one by one through biological experiments under the constraints of time and cost. In addition, the results of the experimental methods will be accompanied with higher false positives and false negatives due to various external factors [6]. Benefiting from the development of high-throughput technologies such as Microarray, Q-PCR, and yeast two-hybrid screens (Y2H) [7, 8], construction of association prediction framework that provide a new viewpoint for gaining a holistic understanding in different fields will be possible based on the published online database such as LncRNADisease [9], HMDD [10], and STRING [11].

In recent years, several computational methods based on public data sets have been put forward successively and was carried on in practice to guide and support manual experiments [12]. These proposed methods can be roughly divided according to the research field, calculation model, calculation method, etc. The prediction model can be divided into several categories because of the different research objects and the typical representative is as follows. In the field of protein–protein interaction (PPI), Wang *et al.* regarded the protein sequence as a kind of natural language called Bio2Vec for feature extraction and discover the potential association by convolution neural network (CNN) [13]. In the field of ncRNA-protein (RPI), Yi *et al.* proposed a robust deep learning framework for predicting interactions through evolutionary information [14].

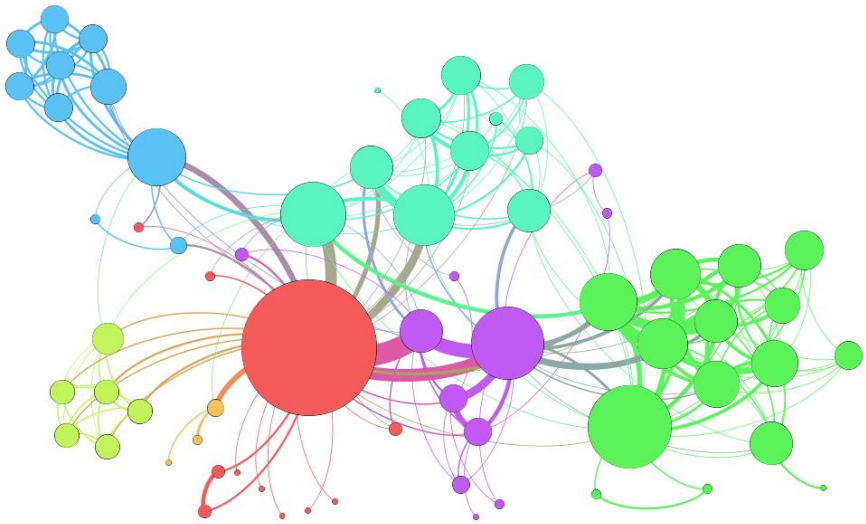

Figure 1. Example of the molecular associations network. Different color represents different molecules nodes and associations edges.

In the field of drug-target interactions (DTI), Wang *et al.* predict the association between drug and target by Rotation Forest based on drug structure and protein sequence [15]. Through the different computational models, the prediction framework can be roughly divided into machine learning based methods, network based methods and matrix decomposition based methods. You *et al.* proposed a novel model called PCA-EELM to predict protein-protein interactions by machine learning model with only information of protein sequence [16]. Chen *et al.* presented a network-based framework to predict miRNA-disease association by integrating known associations and the similarity of miRNAs and diseases respectively [17]. Li *et al*. transformed the problem of discovering undetected miRNA-disease association into the problem of adjacency matrix completion and proposed a prediction model called MCMDA [18]. Most of the existing computational models are based on direct associations or the characteristics of the research objects themselves to detect unknown relationships. And now, it is becoming more and more popular to explore potential associations through some intermediary. For example, by constructing a heterogeneous network of miRNA, lncRNA, and disease, Chen *et al.* took lncRNA as an intermediary to discover miRNA-disease associations through label propagation algorithms [19]. Peng *et al.* carried on CNN as the classifier to predict undetected miRNA-disease associations by capturing similarity in a three-layer network including miRNA, protein, and disease [20]. Researchers are gradually addressing this problem through an increasingly overall perspective, but to date there is still no predictive model that can discover the association of any node in the complete network within a cell.

In this study, we present a model to predict the relationship between any small molecules in a cell based on sequence and network embedding through a more systematic and comprehensive view. The complex associations network of biomolecules (as shown in Figure 1 and Figure 2) consists of two parts: nodes (ncRNA (miRNA, lncRNA), protein (target), drug, disease) and edges (the relationship between nodes). Determining the edges between any two nodes in the whole complex

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 July 2019 doi:10.20944/preprints201907.0043.v1

Peer-reviewed version available at Cells 2019, 8, 866; doi:10.3390/cells8080866

network helps people to have a deep and comprehensive understanding of various life activities in living organisms from another micro perspective [21, 22].

Firstly, nine kinds of molecular associations, such as miRNA-disease association, protein-protein interaction, lncRNA-protein interactions, and drug-target interaction were collected to consider the relationship between each node and any other kind of node in a global way. After de-redundancy and repetition, five research objects such as miRNA, protein, and drug were obtained and co-combined to construct a complex heterogeneous network in an entire view at the cellular level. Secondly, each node can be represented in two ways. One is the node intrinsic attributes such as the sequence of ncRNA and protein, the molecular fingerprint of drug, and the phenotype of disease. The other way is to represent the relationship between nodes and other nodes as a vector through network embedding. Thirdly, all known associations are treated as positive samples, and an equal number of unknown associations are randomly selected as negative samples, which together serve as a training set. Random forest is selected as a classifier for training verification and testing. The five-fold cross validation was adopted to evaluated the proposed method, and we also have compared the performance of different types of features, classifiers and previous methods. The results indicate that our method combined intrinsic attribute feature and manner information could achieve effective and robust prediction performance. The construction of systematic and complex molecular associations network offers a new view, which can help us better understand biology and disease pathologies. To the best of our knowledge, we are the first to construct molecular associations network using associations between lncRNA, miRNA, disease, drug and protein. We hope that this work will inspire more research on representational learning on biological networks.

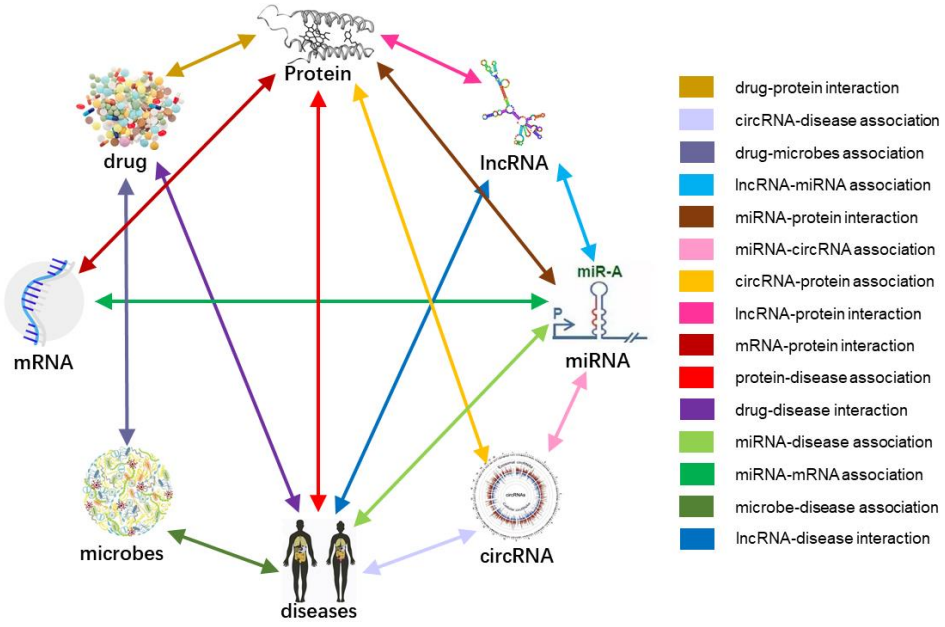

Figure 2. The molecular associations network.

2. Materials and Methods

2.1. Combined nine kind of associations to construct the molecular associations network

In order to systematically and holistically establish a biomolecular relationship network, known associations between biological small molecule transcripts (miRNA, lncRNA and protein), diseases and drugs were downloaded from multiple databases. The details of the final experimental data obtained after performing the inclusion of identifier unification, de-redundancy, simplification and deletion of the irrelevant items are shown in the following Table 1.

Table 1. The details of nine kinds of associations in the molecular associations network.

Relationship type	Database	Number of associations
miRNA-lncRNA	lncRNASNP2 [23]	8374
miRNA-disease	HMDD [10]	16427
miRNA-protein	miRTarBase [24]	4944
lncRNA-disease	LncRNADisease [9], lncRNASNP2 [23]	1264
lncRNA-protein	LncRNA2Target [25]	690
protein-disease	DisGeNET [26]	25087
drug-protein	DrugBank [27]	11107
drug-disease	CTD [28]	18416
protein-protein	STRING [11]	19237
Total	N/A	105546

After aggregating the above database, we separately classify the different nodes to get the final statistics as shown in the following Table 2.

Table 2. The amount of 5 types of nodes in the molecular associations network.

Node	Amount
Disease	2062
LncRNA	769
MiRNA	1023
Protein	1649
Drug	1025
Total	6528

2.2. NcRNA and Protein Sequence

The sequences of miRNA, lncRNA, and protein are downloaded from miRbase [29], NONCODE [30], and STRING [11], respectively, to subsequently represent the attribute of the node. For the sake of simplicity, we chose to encode ncRNA sequences using a 64 (4 × 4 × 4) dimensional vector, in which each feature represents the normalized frequency of the corresponding 3-mer appearing in the RNA sequence (e.g. ACG, CAU, UUG). Inspired by the article of Shen *et al*. [31], in the process of protein sequence encoding, 20 amino acids are classified into 4 classes according to the polarity of the side chain including (Ala, Val, Leu, Ile, Met, Phe, Trp, Pro), (Gly, Ser, Thr, Cys, Asn, Gln, Tyr), (Arg, Lys, His) and (Asp, Glu). Thus, each protein sequence can be represented by a 3-mer that is 64 (4 × 4 × 4) dimensional vector and each dimension of the vector representing the normalized frequency of the corresponding 3-mer in the sequence.

2.3 Disease MeSH Descriptors and Directed Acyclic Graph

The Medical Subject Headings (MeSH) is a comprehensive searchable control vocabulary which is organized by National Library of Medicine furnished a rigorous index for journal articles and books in the life sciences. The top-level categories in the MeSH descriptor hierarchy are: Anatomy [A], Organisms [B], Diseases [C], Chemicals Drugs [D] and so on. In this system, each disease can be represented by a Directed Acyclic Graph (DAG) generated through its MeSH, accurately and objectively describe its own characteristics. The details to describe the disease with DAG is as follows. $DAG(D) = (D, N(D), E(D))$, $N(D)$ is the set of points that contains all the diseases in the $DAG(D)$. $E(D)$ is the set of edges that contains all relationships between nodes in the $DAG(D)$. An example of the disease Astrocytoma's DAG is as follows Figure 3:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 July 2019 doi:10.20944/preprints201907.0043.v1

Peer-reviewed version available at Cells 2019, 8, 866; doi:10.3390/cells8080866

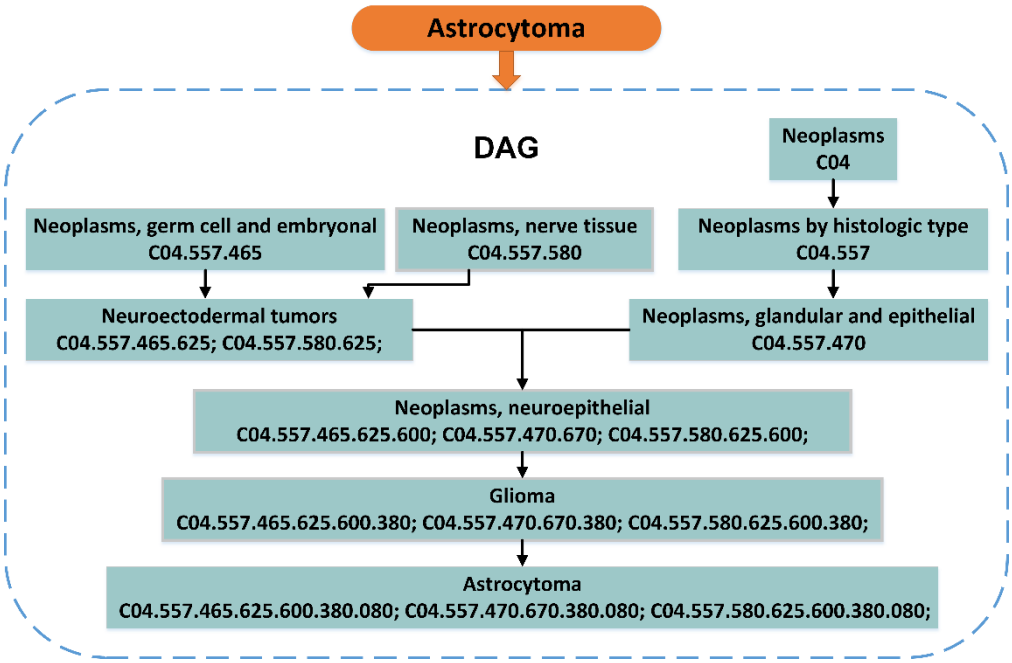

Figure 3. Construction of a disease's DAG.

For the diseases that are included in MeSH, the semantic similarity that is calculated by means of DAG can be chose to represent the disease according to the previous literature [32]. The semantic similarity between different diseases can be defined as follows. In DAG of disease D, the contribution of any ancestral disease t to disease D is as the formula:

$$\begin{cases} D1_D(t) = 1 & if\ t = D \\ D1_D(t) = \max\{\Delta * D1_D(t') | t' \in children\ of\ t\} & if\ t \neq D \end{cases} \quad (1)$$

Δ is the semantic contribution factor. The contribution of disease D to itself is 1 and the contribution of other nodes to D will be attenuated due to Δ. Based on Equation (1), we can obtain the sum of the contributions of all diseases in DAG to D:

$$DV1(\mathrm{D}) = \Sigma_{t \in N_D} D1_D(t) \quad (2)$$

Similar to the Jaccard similarity coefficient, the semantic similarity between the diseases i and j can be calculated by the following formula:

$$S1(i,j) = \frac{\sum_{t \in N_i \cap N_j} (D1_i(t) + D1_j(t))}{DV1(i) + DV1(j)} \quad (3)$$

2.4. Drug Molecular Fingerprint

The smiles of drugs were downloaded from DrugBank and then transformed into corresponding Morgan Fingerprint by python package.

2.5. Stacked Autoencoder

In order to reduce noise and normalize attribute information to a uniform dimension, stacked autoencoder was employed to obtain a suitable subspace from the original feature space. SAE can be divided into two parts: the encoder that encodes the input data into corresponding representation h and the decoder that reconstructs an approximation $\hat{x}$ from the hidden representation h.

$$h = f(x) \coloneqq S_f(Wx + p) \quad (4)$$

$$y = g(h) \coloneqq S_g(W'x + q) \quad (5)$$

Here, we choose the ReLU function as the activation function:

$$S_f(t) = S_g(t) = max(0, Wt + b) \quad (6)$$

2.6. Node Representation

In the entire biomolecular network, each node is represented by its intrinsic attributes and its relationship with other nodes. The attributes of the node itself can be the sequence of ncRNA, protein, the semantics of the disease, and the molecular fingerprint of the drug. The relationship with other nodes can be considered as a functional representation based on the idea of collaborative filtering. More specifically, in this work, we chose a method of network embedding called *LINE* [33] to globally represent the manner of nodes in the entire network and the flow of information directly or latently with other nodes.

For large-scale networks, some existing network representation learning algorithms require complex computational complexity. Recently, some methods of large-scale networks either use indirect methods to reduce computational complexity or lack explicit objective function (DeepWalk [33]). LINE [31] defines two similarity relationships, including the first-order proximity and the second-order proximity. The first-order similarity is defined as the node connection relationship (local feature) in the network, and the second-order similarity is defined as the common neighbor node (global feature) of the nodes that are not directly connected as a supplement to the first-order similarity. In this work, we use the network representation model LINE to learn how to represent the relationships between each node and other nodes in the entire network. In this way, an undirected edge can be considered as two directed edges with opposite directions and equal weights. The second-order proximity assumes that vertices sharing many connections to other vertices are similar to each other. The probability that v_j is a neighbor of v_i is defined as:

$$p_2(v_j|v_i) = \frac{exp(\vec{u}_j^T \cdot \vec{u}_i)}{\sum_{k=1}^{|V|} exp(\vec{u}_j^T \cdot \vec{u}_i)} \quad (6)$$

The probability that each point in the network is a neighbor of vi) is defined as:

$$\hat{P}_2(v_j|v_i) = \frac{w_{ij}}{d_i} \quad (7)$$

Therefore, we minimize the following objective functions:

$$O_2 = \sum_{i \in V} \lambda_i d(\hat{P}_2(\cdot|v_i), p_2(\cdot|v_i)) \quad (8)$$

For the sake of simplicity, λ_i is set to the degree of the vertex *i*, *i.e.* $\lambda_i = d_i$. Here KL divergence is used as the function of distance. After some constants are omitted, the loss function can be simplified as the following form:

$$O_2 = -\sum_{(i,j) \in E} w_{i,j} \log p_2(v_j|v_i)) \quad (9)$$

3. Results and Discussion

3.1. Evaluate the 5-fold cross validation performance of our method

For the five-fold cross-validation, the entire data set was randomly divided into five subsets of equal size, one subset was treated as the test set in turn, and the remaining four subsets were used as the training sets to construct the classifier. Note that at the time of each cross validation, only the currently training set, *i.e.* 80% of the total edges, would be embedding as the manner of the node, which avoids the leakage of test information. Although the above operations may cause some of the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 July 2019 doi:10.20944/preprints201907.0043.v1

Peer-reviewed version available at Cells 2019, 8, 866; doi:10.3390/cells8080866

nodes originally in the network to become isolated *i.e.* degree with 0 and these nodes may also lack attribute information at the same time. This situation can better simulate the real environment to provide support and assistance for the exploration of unknown fields by researchers through manual experiments.

A range of broader evaluation criteria including accuracy (Acc.), sensitivity (Sen.), specificity (Spec.), precision (Prec.) and MCC was utilized to evaluate the proposed model more comprehensively and fairly. As shown in Table 3 and Figure 4, the results of average Acc., Sen., Spec., Prec., MCC and AUC of 92.38%, 92.61%, 92.14%, 92.18%, 84.76% and 97.35% when the proposed framework was applied to predict arbitrary associations in the whole network. The details of 5-fold cross-validation results performed by our method were list in Table 1. Receiver operating characteristic curve (ROC) is a commonly used standard for evaluating models. Area under curve (AUC) is the area of graph which is surround by the roc, the abscissa false positive rate (FPR), and the ordinate true positive rate (TPR). We also draw the ROC and calculated AUC to visually evaluate our proposed model at the same time as the 5-cross validation. PR curve whose abscissa is recall and ordinate is precision was applied to evaluated the model from another angle. In conclusion, our method obtained AUC of 0.9735 and AUPR (area under PR) which indicated that the proposed method combined 2 kinds of information had excellent ability to identify positive and negative samples. The higher AUC and AUPR indicated our method had a strong predictive performance and the lower variance of the results showed the proposed model was stable and robust.

Table 3. 5-Fold cross-validation results performed by our method.

fold	Acc. (%)	Sen. (%)	Spec. (%)	Prec. (%)	MCC (%)	AUC (%)
0	92.25	92.68	91.82	91.89	84.51	97.35
1	92.43	92.52	92.35	92.36	84.87	97.34
2	92.49	92.84	92.13	92.19	84.98	97.39
3	92.58	92.75	92.42	92.44	85.16	97.39
4	92.13	92.28	91.98	92.01	84.26	97.29
Average	**92.38±0.18**	**92.61±0.22**	**92.14±0.25**	**92.18±0.23**	**84.76±0.37**	**97.35±0.04**

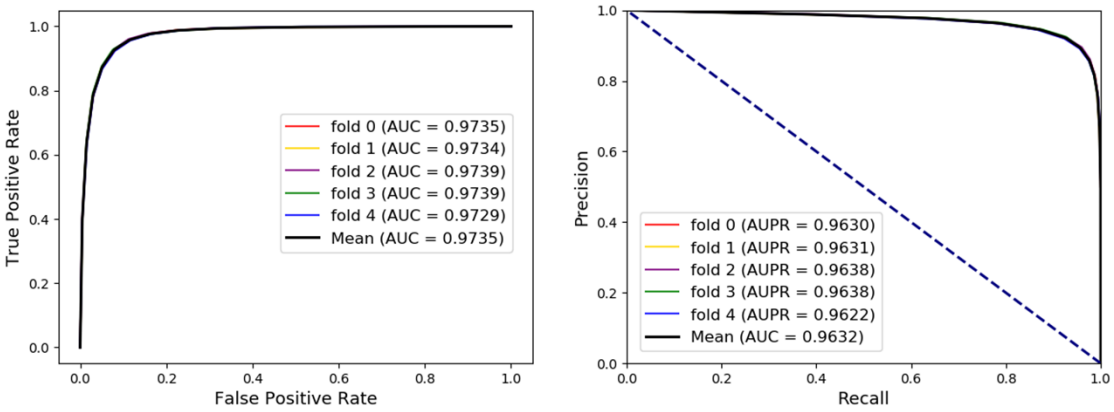

Figure 4. The ROCs, AUCs, PRs, and AUPRs of our method under 5-fold cross validation on the whole dataset.

3.2 Comparison of different feature extraction methods

As mentioned above, each node in the biomolecular network within cell can be represented by two kinds of information including attribute information and manner information. In order to evaluate the impact of each type of information on the final classification effect, we respectively utilized the information of attribute, the information of manner and the combination of the above two to represent the node under the extensive evaluation criteria in the 5-fold cross-validation. As shown in Table 4 and Figure 5, the average of ROC and PR under 5-fold cross validation is reported.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 July 2019 doi:10.20944/preprints201907.0043.v1

Peer-reviewed version available at Cells 2019, 8, 866; doi:10.3390/cells8080866

A variety of evaluation criteria as shown in the table below indicated that the node representation combined with the two kinds of information has more outstanding expressiveness.

Table 4. Comparison of different features.

feature	Acc. (%)	Sen. (%)	Spec. (%)	Prec. (%)	MCC (%)	AUC (%)
Attribute	88.62±0.14	91.48±0.13	85.76±0.2	86.53±0.17	77.37±0.28	94.47±0.11
Manner	90.7±0.14	88.84±0.15	92.56±0.19	92.27±0.19	81.45±0.29	96.26±0.05
Both	**92.38±0.18**	**92.61±0.22**	**92.14±0.25**	**92.18±0.23**	**84.76±0.37**	**97.35±0.04**

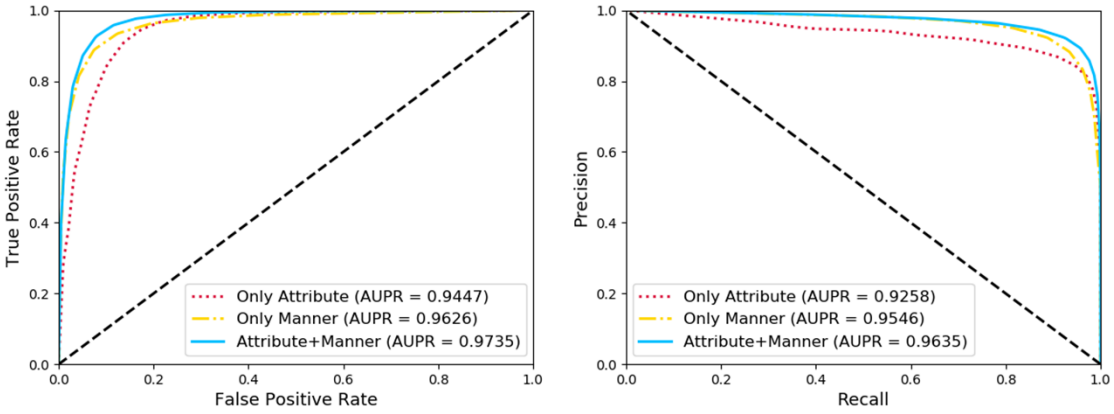

Figure 5. Comparison with different features under 5-fold cross validation.

3.3 Comparison of different classifers

In order to evaluate the performance of the classifier, we compared Random Forest with Adaboost, Logistic Regression, Naïve Bayes and XGBoost under 5 cross-validation in various evaluation criteria. Under the control variable method, the various settings of the experiment are the same except the classifier which makes the comparison of experimental results fairer and more credible. The results are shown in Table 5 and Figure 6. The difference in the effect of different classifiers may be caused by the following factors: (1) Naive Bayes can get better results when the properties of the sample are independent of each other. In this experiment, there are cases where the attributes are not independent and cross-joining together affects the final classification effect. (2) Logistic Regression is essentially a linear classifier whose performance is limited by the distribution of the data and did not perform well in this article. (3) The parameters of all classifiers are default values, which may cause Adaboost and XGBoost to have under-fitting or over-fitting on this task.

Table 5. Comparison of different classifiers.

Classifier	Acc. (%)	Sen. (%)	Spec. (%)	Prec. (%)	MCC (%)	AUC (%)
Adaboost	80.03±0.29	80.91±0.3	79.14±0.43	79.51±0.36	60.07±0.58	87.99±0.28
Logistic	79.92±0.29	82.78±0.29	77.06±0.49	78.3±0.37	59.94±0.57	87.47±0.26
Naive Bayes	55.93±0.15	24.83±0.24	87.04±0.32	65.7±0.5	15.15±0.41	72.13±0.34
XGBoost	84.37±1.3	82.89±2.96	85.85±0.56	85.42±0.37	68.8±2.58	92.7±0.66
Random Forest	**92.38±0.18**	**92.61±0.22**	**92.14±0.25**	**92.18±0.23**	**84.76±0.37**	**97.35±0.04**

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 July 2019 doi:10.20944/preprints201907.0043.v1

Peer-reviewed version available at Cells 2019, 8, 866; doi:10.3390/cells8080866

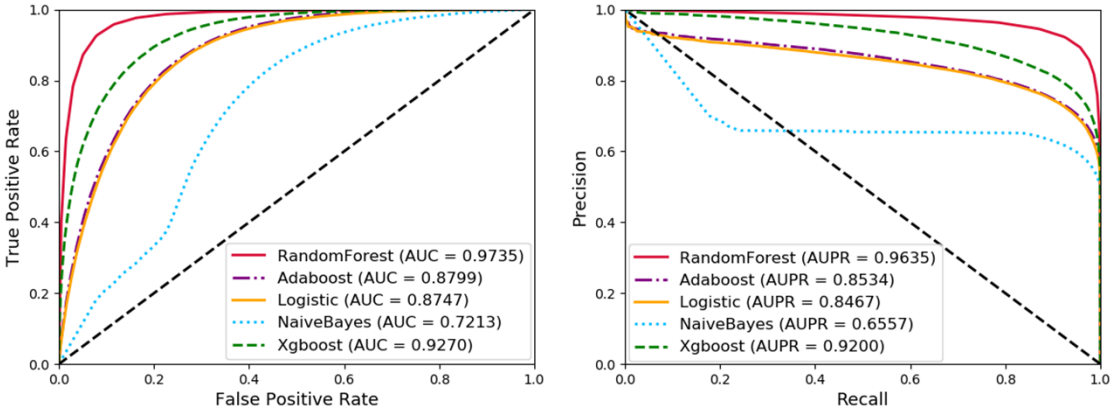

Figure 6. Comparison with Random Forest, Adaboost, Logistic Regression, Naive Bayes and XGBoost under 5-fold cross validation.

3.4 Additional comparision experiment for lncRNA-disease association prediction

In order to compare with traditional methods that focus on single or isolated objects, the lncRNA-disease association prediction was chosen to perform this comparison experiment because of the serious lack of node attribute information. After processing the data, 1263 independent lncRNA-disease association pairs were obtained including 345 different lncRNAs and 295 different diseases.

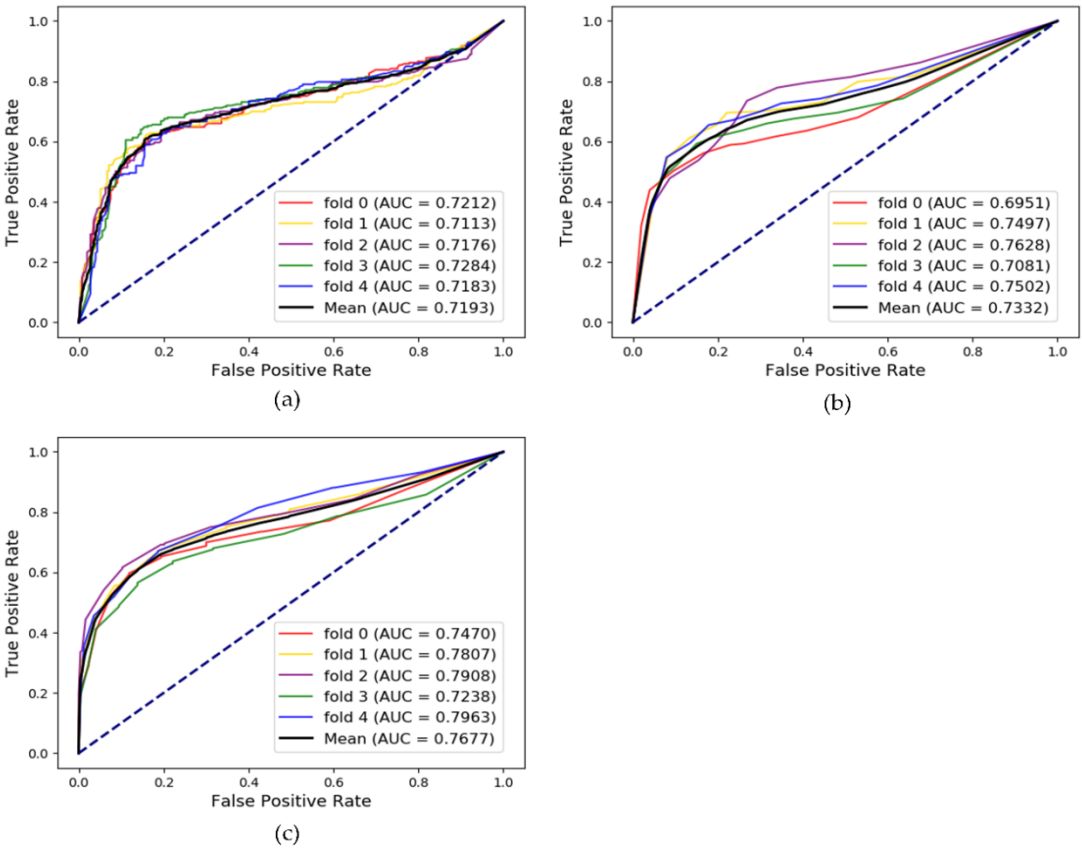

Figure 7. Comparison with previous methods for lncRNA-disease association prediction.

The sequence of each lncRNA was determined when the experimental material was collected. However, among all diseases associated with lncRNA, only 76 of 295 diseases were able to obtain attribute information by constructing DAG to produce similarity with other diseases. The pairs which include both 2 kind of information only take possession of 259 in 1263 associations. Figure 7a showed results of the link prediction 5-fold cross validation with pure attribute information as the characteristics of the node. Figure 7b showed the results of link prediction based on the feature combined attribute information with previous isolated embedding method.

That is, 80% lncRNA-disease associations were utilized to construct the adjacency matrix for generating Gaussian Profile Kernel Similarity in each fold cross validation [34]. Figure 7c shows the results of global embedding, that is, 80% of the lncRNA-disease associations and all the other 8 kind of relationships were processed by LINE in each cross-validation. Obviously, after combining the global manner information, the performance of prediction in lncRNA-disease association can be greatly improved. It also proves that the cell is a complete unit of life, and the interaction of biomolecules in the cell together maintains the normal conduct of life activities.

4. Conclusions

Accumulating evidence demonstrates the superiority of link prediction based on massive data through machine learning models, which not only serves as an addition to manual experiments, but also provides researchers with an overall and macro insight into the interactions between intracellular molecules. In this article, we proposed a new framework based on 5 different kind of nodes and 9 different kind of relationships to detect any potential associations between arbitrary research objects in the whole network. Each node can be represented as a vector by two kinds of information including node attributes and node manner. For attribute information, ncRNA and protein could be encoded into 64-dimensional vectors by the method of k-mer, in which each feature represents the normalized frequency of the corresponding 3-mer appearing in the RNA or protein sequence. The characteristics of the disease and the drug can be represented by their own semantic and molecular structure and transformed into 64-dimensional vectors through the function of feature selection and transformation in SAE. For manner information, the relationship of each node with others could be abstracted by the network embedding method LINE. Combined with the above two kinds of information, each node can be represented by a 128-dimensional vector and put into Random Forest to carry out prediction. The experimental results provide that our method can achieve outstanding performance. The construction of molecular regulatory network in human cells, offer a new systematic view on understanding complex life activities and diseases.

Author Contributions: conceptualization, Z-H.G. and Z-H.Y.; methodology, H-C.Y.; software, investigation, resources and data curation, Z-H.G.; writing—original draft preparation, Z-H.G. and H-C.Y.; validation, visualization and formal analysis, H-C.Y.; writing—review and editing, Z-H.Y. and H-C.Y.; project administration, funding acquisition and supervision, Z-H.Y.;

Funding: This research was funded by the National Natural Science Foundation of China, grant number 61772333.

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

References

1. Ambros V: MicroRNA pathways in flies and worms: growth, death, fat, stress, and timing. Cell 2003, 113(6):673-676.
2. Ponting CP, Oliver PL, Reik W: Evolution and functions of long noncoding RNAs. Cell 2009, 136(4):629-641.
3. Bonetta L: Protein–protein interactions: interactome under construction. Nature 2010, 468(7325):851.
4. Skrabanek L, Saini HK, Bader GD, Enright AJ: Computational prediction of protein–protein interactions. Molecular biotechnology 2008, 38(1):1-17.
5. Ashburn TT, Thor KB: Drug repositioning: identifying and developing new uses for existing drugs. Nature reviews Drug discovery 2004, 3(8):673.

6. Collins SR, Kemmeren P, Zhao X-C, Greenblatt JF, Spencer F, Holstege FC, Weissman JS, Krogan NJ: Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. Molecular & Cellular Proteomics 2007, 6(3):439-450.
7. Yan B, Wang Z-H, Guo J-T: The research strategies for probing the function of long noncoding RNAs. Genomics 2012, 99(2):76-80.
8. Fields S, Song O-k: A novel genetic system to detect protein–protein interactions. Nature 1989, 340(6230):245.
9. Chen G, Wang Z, Wang D, Qiu C, Liu M, Chen X, Zhang Q, Yan G, Cui Q: LncRNADisease: a database for long-non-coding RNA-associated diseases. Nucleic acids research 2012, 41(D1):D983-D986.
10. Huang Z, Shi J, Gao Y, Cui C, Zhang S, Li J, Zhou Y, Cui Q: HMDD v3. 0: a database for experimentally supported human microRNA–disease associations. Nucleic acids research 2018, 47(D1):D1013-D1017.
11. Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, Santos A, Doncheva NT, Roth A, Bork P: The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. Nucleic acids research 2016:gkw937.
12. Yi H-C, You Z-H, Zhou X, Cheng L, Li X, Jiang T-H, Chen Z-H: ACP-DL: A Deep Learning Long Short-Term Memory Model to Predict Anticancer Peptides Using High Efficiency Feature Representation. Molecular Therapy - Nucleic Acids 2019.
13. 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 2019, 8(2):122.
14. Yi H-C, You Z-H, Huang D-S, Li X, Jiang T-H, Li L-P: A Deep Learning Framework for Robust and Accurate Prediction of ncRNA-Protein Interactions Using Evolutionary Information. Molecular Therapy-Nucleic Acids 2018, 11:337-344.
15. Wang L, You Z-H, Chen X, Yan X, Liu G, Zhang W: Rfdt: A rotation forest-based predictor for predicting drug-target interactions using drug structure and protein sequence information. Current Protein and Peptide Science 2018, 19(5):445-454.
16. You Z-H, Lei Y-K, Zhu L, Xia J, Wang B: Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis. In: BMC bioinformatics: 2013: BioMed Central; 2013: S10.
17. Chen X, Xie D, Wang L, Zhao Q, You Z-H, Liu H: BNPMDA: bipartite network projection for MiRNA–disease association prediction. Bioinformatics 2018, 34(18):3178-3186.
18. Li J-Q, Rong Z-H, Chen X, Yan G-Y, You Z-H: MCMDA: Matrix completion for MiRNA-disease association prediction. Oncotarget 2017, 8(13):21187.
19. Chen X, Zhang D-H, You Z-H: A heterogeneous label propagation approach to explore the potential associations between miRNA and disease. Journal of translational medicine 2018, 16(1):348.
20. Peng J, Hui W, Li Q, Chen B, Jiang Q, Wei Z, Shang X: A learning-based framework for miRNA-disease association prediction using neural networks. bioRxiv 2018:276048.
21. Hrdlickova B, de Almeida RC, Borek Z, Withoff S: Genetic variation in the non-coding genome: Involvement of micro-RNAs and long non-coding RNAs in disease. Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease 2014, 1842(10):1910-1922.
22. Barabási A-L, Oltvai ZN: Network biology: understanding the cell's functional organization. Nature Reviews Genetics 2004, 5(2):101-113.
23. 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 research 2017, 46(D1):D276-D280.
24. Chou C-H, Shrestha S, Yang C-D, Chang N-W, Lin Y-L, Liao K-W, Huang W-C, Sun T-H, Tu S-J, Lee W-H: miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. Nucleic acids research 2017, 46(D1):D296-D302.
25. Cheng L, Wang P, Tian R, Wang S, Guo Q, Luo M, Zhou W, Liu G, Jiang H, Jiang Q: LncRNA2Target v2. 0: a comprehensive database for target genes of lncRNAs in human and mouse. Nucleic acids research 2018, 47(D1):D140-D144.
26. Piñero J, Bravo À, Queralt-Rosinach N, Gutiérrez-Sacristán A, Deu-Pons J, Centeno E, García-García J, Sanz F, Furlong LI: DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic acids research 2016:gkw943.
27. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z: DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic acids research 2017, 46(D1):D1074-D1082.

28. Davis AP, Grondin CJ, Johnson RJ, Sciaky D, McMorran R, Wiegers J, Wiegers TC, Mattingly CJ: The comparative toxicogenomics database: Update 2019. Nucleic acids research 2018, 47(D1):D948-D954.
29. Kozomara A, Birgaoanu M, Griffiths-Jones S: miRBase: from microRNA sequences to function. Nucleic acids research 2018, 47(D1):D155-D162.
30. Fang S, Zhang L, Guo J, Niu Y, Wu Y, Li H, Zhao L, Li X, Teng X, Sun X: NONCODEV5: a comprehensive annotation database for long non-coding RNAs. Nucleic acids research 2017, 46(D1):D308-D314.
31. Shen J, Zhang J, Luo X, Zhu W, Yu K, Chen K, Li Y, Jiang H: Predicting protein–protein interactions based only on sequences information. Proceedings of the National Academy of Sciences 2007, 104(11):4337-4341.
32. Wang D, Wang J, Lu M, Song F, Cui Q: Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. Bioinformatics 2010, 26(13):1644-1650.
33. Tang J, Qu M, Wang M, Zhang M, Yan J, Mei Q: Line: Large-scale information network embedding. In: Proceedings of the 24th international conference on world wide web: 2015: International World Wide Web Conferences Steering Committee; 2015: 1067-1077.
34. van Laarhoven T, Nabuurs SB, Marchiori E: Gaussian interaction profile kernels for predicting drug–target interaction. Bioinformatics 2011, 27(21):3036-3043.