Integrative Construction and Analysis of Molecular Association Network in Human Cells by Fusing Node Attribute and Behavior Information

Zhen-Hao Guo,1,2,3 Zhu-Hong You,1,2,3 and Hai-Cheng Yi1,2

Detecting whether a pair of biomolecules associate is of great significance in the study of molecular biology. Hence, computational methods are urgently needed as guidance for practice. However, most of the previous prediction models influenced by reductionism focused on isolated research objects, which have their own inherent defects. Inspired by holism, a machine-learning-based framework called MAN-node2vec is proposed to predict multi-type relationships in the molecular association network (MAN). Specifically, we constructed a large-scale MAN composed of 1,023 miRNAs, 1,649 proteins, 769 long non-coding RNAs (lncRNAs), 1,025 drugs, and 2,062 diseases. Then, each biomolecule in MAN can be represented as a vector by its attribute learned by k-mer, etc. and its behavior learned by node2vec. Finally, the random forest classifier is applied to carry out the relationship prediction task. The proposed model achieved a reliable performance with 0.9677 areas under the curve (AUCs) and 0.9562 areas under the precision curve (AUPRs) under 5-fold cross-validation. Also, additional experiments proved that the proposed global model shows more competitive performance than the traditional local method. All of these provided a systematic insight for understanding the synergistic interactions between various molecules and diseases. It is anticipated that this work can bring beneficial inspiration and advance to related systems biology and biomedical research.

INTRODUCTION

Benefiting from the development of increasingly sophisticated high-throughput technologies, numerous biomolecular relationship networks, such as the non-coding RNA (ncRNA) target-regulation network,1,2 protein-protein interaction network,3,4 ncRNA disease-association network,5,6 etc., are continuously being confirmed to play key roles in cell life activities and processes.7,8 The technical threshold and cost of practical experiments has become affordable and begun to change people’s grasp of biological processes like cell-cycle9 differentiation10 and apoptosis11 from a microscopic perspective. Despite the impressive wet experiment methods making remarkable achievements in sequence identification, relationship mapping, etc., the awareness of relationships between biomarkers remains incomplete. Nevertheless, the approach without guidance is aimless and costly. In addition, identifying relationships between biomolecules by manual experiment is uncertain; especially errors such as false positive rate (FPR) and false negative rate (FNR) can lead to perceived deviations.12 Compared to these practical methods, computational models can learn intrinsic characteristics of biomolecules and infer the potential relationships simultaneously. The accumulation of data and the demand of reality make them become popular. Chen et al.13 utilized stacking automatic encoders for data preprocessing and support vector machines for classification to discover potential microRNA (miRNA)-disease associations. Guo et al.14 predict uncovered long non-coding RNA (lncRNA)-disease associations by combining known associations and disease characteristics. Cheng et al.15 infer new targets for known drugs only through drug-target bipartite network topology similarity. All of the above methods are supporters of reductionism and are the products of compromise under the condition of missing data. They are typical representatives of building, analyzing, and predicting based on a single relationship. Hence, more and more researchers are paying attention to this issue and are working to improve this situation through different strategies. For instance, Lin et al.16 predicted that lncRNA-disease associations were only intermediated by indirect miRNAs and still achieved higher areas under the curve (AUCs) in the leave-one-out cross-validation. Although reductionism provides a wealth of knowledge over a period of time, it ignores that the cell itself is as a whole based on the genetic central dogma. Existing evidence intensely indicates that the function of cells is rarely directly controlled by a determined gene but rather reflects the result of interaction by multiple factors.17 However,
limited by existing incomplete data, computational models are often
directed by reductionism which describes the composition and func-
tion of cells into various parts. Although the circumstance of data loss
is alleviated through different methods, it cannot theoretically corre-

cpond with the genetic central dogma to establish a gene-to-expres-
sion description to explain why there is a relationship between the
biomolecules.

In fact, cells are living organisms with abundant functions under the
relationship of different biology molecules. These different biomole-
cules and their relationships can be treated as vertexes (nodes) and links
(edges) in a network or graph (cell). Graph is an important form of data
that appears widely in the real world and is studied in depth. Since the
"scale-free" and "small-world" network theories were proposed, graphs
have become a research hotspot. Analysis of graphs helps not only
to understand the hidden knowledge behind data, but also to expand
and migrate to other types of data. Network representation is an effec-
tive way to solve this problem. The previous representation algorithm
aims at acquiring the main components to obtain dimensional reduc-
tion, such as singular value decomposition (SVD) and locally linear
embedding (LLE). The maturity of deep-learning technology has
promoted the development of many fields. A large number of new
network representation technologies, such as DeepWalk, node2vec,
and LINE, can more effectively extract the structure of the network
and facilitate downstream tasks such as link prediction, node classifi-
cation, community discovery, and visualization. Inspired by Guo
et al., rescanning some fundamental biological problems from a

global network viewpoint can aid researchers in treating the problem
from a different perspective in order to find a new solution. The differ-
ces between diverse methods can be seen in Figure 1.

In this paper, a relatively complete molecular association network
(MAN) is constructed, including various sub-networks to reveal the
flow of genetic information. The entire network can be described as
follows: protein as a direct participant in the expression of genetic in-
formation is regarded as the core of the entire network. With the
introduction of competing endogenous RNA (ceRNA), the establish-
ment of a link between ncRNA including miRNA, IncRNA, etc.,
proves that they are not transcribed garbage but more or less regu-
lated gene expression. The increasing evidence about the inextricable
associations between ncRNA and disease also confirms the above hy-
pothesis. Relative to the composition inside the cell, factors outside
the cell will have a crucial impact on life activities from another
perspective; the drug-target-disease subnet has long been experiment-
ally proven to play an irreplaceable role in drug development and
reposition. Although intuitively, there are many different types of no-
des and complex interlaced edges in the network, the exchange of in-
formation between different molecules is very clear even under the
overlap of different modules and subnets. The perspective of MAN
is shown in Figure 2.

The relationship prediction problem in the above description network
can be formally defined as follows: A graph is \( G = (V, E) \), where \( V \) and \( E \) are

each element \( A_{ij} \) is equal to 1 if and only if node \( i \) and node \( j \) are
 experimentally verified to be associated. The aim of the model is
to find the uncovered element 1 in \( A \). The main steps of the entire
model are as follows: first, the relationships collected by various data-
bases are summarized after redundancy removal and identifier uni-
form to construct the MAN. After the complex network consisting
of five kinds of nodes and nine kinds of edges can be defined as a ho-

mogeneous undirected graph, adjacency matrix \( A \) can be constructed
to contain all information of nodes and edges. In order to simplify the
calculation and facilitate storage, we only take the lower triangular
part of matrix \( A \). Second, k-mer, node2vec, etc., which are widely
used in bioinformatics and network embedding, are applied to map
the nodes to a low-dimensional dense feature space. The attribute in-
formation and behavior information of each node can be represented
as a 64-dimensional vector, respectively. This process can be called
biomolecular digitization, or biomolecule2vec. Third, relationships
that have been validated by manual experiments are considered pos-
itive samples. An equal number of negative samples from unknown
pairs are randomly extracted. All of the positive and negative samples
are sent to random forest for training and prediction. For fairness, all
parameters are set to default values in each step. There is no need to
guarantee that the degree of each node is bigger than 0 when the

---

**Figure 1. The Differences between Diverse Methods**

(A–C) Comparison between traditional method (A), intermediary-based method (B), and the proposed method (C). The traditional method (A) often focuses on the research itself and ignores the mediation role of other kinds of biomolecules in the cell. The intermediary-based method (B) is often limited without considering second- or higher-order neighbors. The proposed method (C) can reflect the relationships in the cell macroscopically and completely and effectively promote the prediction task.

**Figure 2. The Structure of the MAN**

The relationship prediction problem in the above description network can be formally defined as follows: A graph is \( G = (V, E) \), where \( V \) and \( E \) are

each element \( A_{ij} \) is equal to 1 if and only if node \( i \) and node \( j \) are
 experimentally verified to be associated. The aim of the model is
to find the uncovered element 1 in \( A \). The main steps of the entire
model are as follows: first, the relationships collected by various data-
bases are summarized after redundancy removal and identifier uni-
form to construct the MAN. After the complex network consisting
of five kinds of nodes and nine kinds of edges can be defined as a ho-

mogeneous undirected graph, adjacency matrix \( A \) can be constructed
to contain all information of nodes and edges. In order to simplify the
calculation and facilitate storage, we only take the lower triangular
part of matrix \( A \). Second, k-mer, node2vec, etc., which are widely
used in bioinformatics and network embedding, are applied to map
the nodes to a low-dimensional dense feature space. The attribute in-
formation and behavior information of each node can be represented
as a 64-dimensional vector, respectively. This process can be called
biomolecular digitization, or biomolecule2vec. Third, relationships
that have been validated by manual experiments are considered pos-
itive samples. An equal number of negative samples from unknown
pairs are randomly extracted. All of the positive and negative samples
are sent to random forest for training and prediction. For fairness, all
parameters are set to default values in each step. There is no need to
guarantee that the degree of each node is bigger than 0 when the

---

Molecular Therapy: Nucleic Acids Vol. 19 March 2020 499
The ROC is a curve that determines the abscissa True Positive Rate (TPR) and the ordinate False Positive Rate (FPR) by using the predicted probability of the test sample as the classification threshold. The area enclosed by the ROC and the coordinate axis is called the AUC. The drawing process of the PR curve is similar to the ROC and the area enclosed by the abscissa recall and the ordinate AUPR. In order to fairly and comprehensively evaluate the proposed model, extensive evaluation criteria, including accuracy (acc.), sensitivity (sen.), specificity (spec.), precision (prec.), and matthews correlation coefficient (MCC) are applied from different perspectives. Although the dataset is balanced, we still hope to provide a reference for subsequent models through this overall measurement system. The details of results under 5-fold cross-validation are shown in Table 1 and Figure 4.

Analysis of Figure 4 and Table 1 concludes that the method based on random forest produces satisfactory results on the MAN. The outstanding results of AUC, AUPR, and various evaluation criteria at each fold suggest superior predictive ability of the proposed model, while the lower standard deviation demonstrates the stability and robustness of the prediction method.

**Feature Importance Comparison**

As mentioned above, each node in the biomolecular network can be represented by two kinds of information. It is obvious that there still exists some predictive ability with the model when either kind of information is lost, so the common situation, such as new sample problem, can be alleviated to some extent. In this chapter, we hope to explore the impact of different information on the prediction effect, that is, the practical application value of the proposed model even under the characterization of a single kind of information. In order to verify only the impact of different features on the prediction results, the random forest classifier is set as the default parameter. The ROC, AUC, PR, AUPR, and extensive evaluation criteria under the 5-fold cross-validation are shown in Table 2 and Figure 5.

Feature comparison experiments indicate that attribute and behavior complement each other and contribute to detect potential associations. Although the combination of the two kinds of information shows the best results, even in the case of a single kind of feature, the proposed model is an excellent method that can adapt to various real environments as an auxiliary tool for manual experiments.

**Comparison with Different Classifiers**

Although many classic machine-learning algorithms have achieved great success and impressive influence in both industry and academia, the prediction effect of the traditional algorithm on the dataset of this article is quite different. In this chapter, we compare the performance of several common classifiers, including random forest, Xgboost, Adaboost, logistic regression, and naive Bayes, and try to

---

**Table 1. Performance of the Proposed Model on Various Evaluation Criteria under 5-Fold Cross-Validation**

| Fold | Acc. (%) | Sen. (%) | Spec. (%) | Prec. (%) | MCC (%) | AUC (%) |
|------|----------|----------|-----------|-----------|---------|---------|
| 0    | 91.69    | 90.98    | 92.40     | 92.29     | 83.39   | 96.83   |
| 1    | 91.85    | 90.17    | 92.64     | 92.52     | 83.71   | 96.86   |
| 2    | 91.49    | 90.58    | 92.41     | 92.27     | 83.00   | 96.68   |
| 3    | 91.62    | 90.99    | 92.25     | 92.15     | 83.24   | 96.74   |
| 4    | 91.52    | 90.74    | 92.30     | 92.17     | 83.04   | 96.74   |
| Average | 91.63 ± | 90.87 ± | 92.4 ± | 92.28 ± | 83.28 ± | 96.77 ± |

---
analyze the reasons for this situation. In order to fairly compare the performance of the classifier on this dataset, all parameters are set to default values. The detailed results under 5-fold cross-validation based on different classifiers are as shown in Table 3 and Figure 6.

The results can be explained as follows: (1) For naive Bayes, there might be strong correlations in each dimension of the representation vectors, making the performance unsatisfactory. (2) For logistic regression, the high complexity of the dataset may not be concomitant with the linear classification surface, making it difficult for the logical return to fit the sample. (3) It is curious that random forest shows better classification results than Xgboost and Adaboost with advanced assemble strategies. The reasons for this are probably attributed to the setting of the default parameters making it hard for the latter to fit the data.

**Additional Comparison Experiment Based on lncRNA-miRNA Interaction Prediction**

Predicting multi-type relationships between different biomolecules to evaluate the performance of the model is limited in some respects. The proposed model can predict not only multiple relationships, but also single associations. Considering the large accumulation of ceRNA evidence and miRNA and IncRNA as a hotspot in the field, the IncRNA-miRNA interaction was chosen as a special additional experiment to compare the proposed method with the state-of-the-art model. 8374 IncRNA-miRNA interaction pairs containing 467 different IncRNAs and 254 different miRNAs were downloaded from IncRNASNP2 on April 26, 2019 after removing redundancy and unfirming identifiers. Four different kinds of experiments under 5-fold cross-validation were performed separately, and the results are as shown in Figure 7.

For Figure 7A, each IncRNA or miRNA is represented as a 64-dimensional vector only based on its attribute feature. Thus, each IncRNA-miRNA interaction pair can be viewed as a 128-dimensional vector with a label of 0 or 1. It can be treated as a baseline compared with other methods. For Figure 7B, influenced by the method proposed by Chen, we completely ignore the direct interactions between IncRNA and miRNA and use the remaining eight kinds of relationships in the network to represent IncRNA or miRNA by node2vec. Each IncRNA or miRNA can be represented as a 64-dimensional vector only based on its behavior information. Thus, each IncRNA-miRNA interaction pair can be viewed as a 128-dimensional vector. Figure 7C, the traditional method of measuring functional similarity through the Gaussian profile kernel function proposed by van Laarhoven et al. is essentially a description of single association and is widely used in ncRNA-disease association prediction. Each IncRNA or miRNA can be represented as a 64-dimensional vector based on both attribute and behavior information. The results indicate that this classic state-of-the-art method has a positive effect on the discovery of potential interactions. For Figure 7D, this is the performance of the proposed method. Each node is represented by both attribute and behavior information. When node2vec is implied, 80% IncRNA-miRNA interactions and the other eight kinds of associations together describe the behavior of the node. The outstanding performance of the proposed method can be attributed to two aspects: first, node2vec is a more advanced algorithm that can extract structural information from the network more easily than the Gaussian profile kernel function. Second, MAN as a whole contains more abundant biological information than direct IncRNA-miRNA interactions. All above experiments show that the additional relationships in MAN indeed contain an amount of biology information and can be used as a kind of auxiliary information to assist the prediction of a single research object.

**DISCUSSION**

Networks naturally exist in a wide diversity of real-world scenarios, e.g., social networks, citation networks, knowledge networks, etc. Effective network analytics provides researchers a deeper understanding of what is behind the data and provides insights into how to make good use of this information.

Inspired by holism, we constructed a MAN by integrating different types of biomolecules to analyze and describe the state and function of various modules from different angles. The computational model called MAN-node2vec was proposed based on MAN to predict arbitrary relationships between any nodes from a global perspective and achieved remarkable prediction performance. Additional experiments indicate that even on specific issues, the proposed approach demonstrated a more competitive ability than traditional methods. All the results demonstrate the feasibility and superiority of uncovering...
potential associations from a global perspective. In fact, holism is not a negation of reductionism, but complementary.

Generally, our research will expand the research paradigm of computational biology and establish interesting connections between biological data and complex network techniques. It can be seen as a foundation for advancing both methodology and technology. MAN-node2vec is not only a supplement to manual experiments, but also a new chapter on the study of the laws of life sciences based on collection, integration, and data mining from a comprehensive perspective.

MATERIALS AND METHODS

Construction of the MAN

Since the existing database does not integrate all the data we need, we have to collect diverse associations from multiple databases as the basis of construction for the MAN.29–37 Specifically, nine kinds of associations that are the edges or links in MAN can be obtained from the corresponding databases shown in Table 4.

The identifiers of miRNA, lncRNA, protein, and drug are based on the nomenclature provided by miRBase, NONCODE, STRING, and DrugBank, respectively. The identifiers of disease are used directly from the original database. After the operation of identifier transformation, redundancy removal, and data filtering from different databases, we get a total of 6,528 biomolecules (nodes) including five different types. The statistics on details of diverse nodes are shown in Table 5.

Obviously, 105,546 experimental valid association pairs can be treated as positive samples. The remaining unlabeled samples are composed of true negative samples and potential positive samples.

ncRNA and Protein Sequence

We collected the sequences of miRNA, lncRNA, and protein from miRbase,38 NONCODE,39 and String,37 respectively, and processed them as described in Shen et al.40 It is well known that the RNA sequence is composed of four kinds of nucleotides: A, adenine; G, guanine; C, cytosine; and U, uracil. The protein sequence is composed of 20 kinds of amino acids, which is quite unfriendly for encoding and storage. Therefore, we divide the 20 kinds of amino acids into four groups, including (1) Ala, Val, Leu, Ile, Met, Phe, Trp, Pro; (2) Gly, Ser, Thr, Cys, Asn, Gln, Tyr; (3) Arg, Lys, His; and (4) Asp, Glu, according to the polarity of the side chain. Thus, each sequence of ncRNA or protein can be represented by a vector in which each dimension of the vector can treated as normalized frequency of the k-mer in the sequence.

In this article, k is set to 3 and each sequence is represented as a 64-dimensional vector (4^3 × 4^3). Each dimension of the vector is the full array of three nucleotide combinations of AAA, AAC, ..., UUU. A window of size 3 can get all the fragments of the current sequence when sliding in steps of 1. Each dimension value of the representation vector can be obtained by counting and normalizing the appearance number of these fragments.

Disease MeSH Descriptors and Directed Acyclic Graph

Medical subject headings (MeSHs) is a rigorous term developed and published by the National Library of Medicine for use in management and inquiries in the fields of biology and medicine. The previous work of calculating similarities through the MeSH descriptors to define the disease is effective and attracts widespread attention.41 Thus, we

Table 3. Comparison of Different Classifiers on Various Evaluation Criteria

| Classifier   | Acc. (%) | Sen. (%) | Spec. (%) | Prec. (%) | MCC (%) | AUC (%) |
|--------------|----------|----------|-----------|-----------|---------|---------|
| NaiveBayes   | 60.59 ± 20.58 | 73.98 ± 0.68 | 47.20 ± 0.79 | 58.36 ± 0.47 | 21.98 ± 1.19 | 70.91 ± 0.47 |
| Logistic     | 76.89 ± 0.30  | 78.92 ± 0.19  | 74.86 ± 0.56  | 75.85 ± 0.41  | 53.83 ± 0.58  | 83.75 ± 0.46  |
| AdaBoost     | 77.78 ± 0.18  | 80.03 ± 0.30  | 75.52 ± 0.13  | 76.58 ± 0.14  | 55.61 ± 0.37  | 85.19 ± 0.18  |
| XgBoost      | 86.17 ± 0.31  | 87.39 ± 0.49  | 84.95 ± 0.57  | 85.31 ± 0.46  | 72.37 ± 0.61  | 93.20 ± 0.22  |
| Random forest| 91.63 ± 0.14  | 90.87 ± 0.20  | 92.4 ± 0.15   | 92.28 ± 0.15   | 83.28 ± 0.29  | 96.77 ± 0.07  |

Considering that the potential positive samples are a small part of all unlabeled samples, a common method widely used in bioinformatics that randomly extracts the same number unlabeled samples as negative samples is applied. Finally, the whole sample set is composed of 211,092 association pairs. Data used for analysis are available on GitHub page: https://github.com/CocoGzh/MAN-1.0.
adopted this method and filtered out the disease-related keywords of the data downloaded from https://www.nlm.nih.gov/.

In this system, each disease with a descriptor can generate a directed acyclic graph and can be accurately and comprehensively characterized. The details of the calculation are described below: The directed acyclic graph (DAG) of disease \( D \) is defined as \( \text{DAG}(D) = (D, N(D), E(D)) \), where \( N(D) \) is a set of points containing all diseases in the DAG, and \( E(D) \) is a set of edges containing all relationships between diseases in the DAG. The contribution of the ancestral node to the disease \( D \) in the directed acyclic graph is calculated by the following formula:

\[
D_D(t) = \begin{cases} 
1 & \text{if } t \in \text{ancestors of } D \\
\max\{D(t') + \Delta D(t') | t' \in \text{children of } t\} & \text{if } t \in D 
\end{cases}
\] (Equation 1)

\( t \) is the element of \( N(D) \) and \( \Delta \) is an attenuation factor. The farther the ancestral disease distance \( D \) is, the smaller the contribution to itself and is defined as 1. The total contribution of all elements in the set \( N(D) \) to disease \( D \) is

\[
\text{DV}(D) = \sum_{t \in N(D)} D_D(t).
\] (Equation 2)

The similarity between disease \( i \) and disease \( j \) can be calculated by the following formula

\[
\text{Similarity}(i, j) = \frac{\sum_{t \in N(i)} D_D(t) + D_D(t)}{\text{DV}(i) + \text{DV}(j)}.
\] (Equation 3)

**Drug Morgan Molecular Fingerprint**

RDKit is an open-source chemical informatics and machine learning toolkit. SMILES (simplified molecular input line entry specification), which was proposed by David Weininger, is a specification for clearly describing molecular structure using ASCII strings. The SMILES were downloaded from DrugBank and converted to Morgan molecular fingerprint by the python package called RDkit to represent the characteristics of the drug.

**Sparse Autoencoder**

After obtaining the high-dimensional representation vector from disease semantics and drug Morgan molecular fingerprints, we use the sparse autoencoder to reconstruct new vectors from the original space to improve feature quality and reduce noise. The sparse autoencoder is an unsupervised learning algorithm that uses the backpropagation algorithm to make the output value as equal as possible to the input value. It consists of two parts, including the encoder that performs the compression function and the decoder that performs the reconstruction function. In addition to the input layer, the input to the \( i \)-th node of the \( l \)-th layer is

\[
Z_i = W_{li} a_{li} + b + \text{Relu}(\text{Zi})
\] (Equation 4)

where \( W_{li} \) is the weight of the \((l-1)\)-th layer neuron to the \( i \)-th neuron of the \( l \)-th layer, \( a_{li} \) is the number of neurons in the \((l-1)\)-layer, and \( b \) is the output of the \((l-1)\)-layer of neurons.

The output of the \( i \)-th node of the \( l \)-th layer is

\[
a_i = f(Z_i + b).
\] (Equation 5)

where \( b \) is the bias and \( f \) is the activation function. Relu was chosen to perform this operation.

\[
f(x) = \max(0, x)
\] (Equation 6)

The loss function is defined as follows:

\[
L(W, b) = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{2} ||a^{(m)}(x^{(i)}) - y^{(i)}||^2 + \alpha \sum_{j=1}^{n} KL(\mu || \hat{\mu}) + \beta w^2.
\] (Equation 7)

The first part is like a normal autoencoder, describing the error between input and output, where \( m \) is the number of samples in the training set, and \( n \) is the number of hidden layers. The second part is a sparsity penalty term called Kullback-Leibler (KL) divergence used to constrain the activity of the hidden layer unit, where \( n \) is the number of hidden layer units. The third part is weight decay to help prevent overfitting.

**Node2vec**

The behavior information of a node can also be considered as a measure of the function of this node. A row or column in the adjacency matrix is a one-hot description of such information. Considering the disadvantages of sparseness, discreteness, and occupying a large amount of storage space, we hope to find a kind of simple and efficient low-dimensional representation.

Node2vec is a kind of representation algorithm with the purpose of mapping of nodes to a new low-dimensional feature space and at
the same time maximizing the preservation of the network structure in the original space. The main idea of node2vec is to treat the random walk path of the nodes in the network, that is, the node sequence, as a text, and then use word2vec to model the path, maximize the likelihood probability, and learn the parameters through the random gradient. By introducing two parameters $p$ and $q$, breadth-first search and depth-first search are introduced into the generation process of random-walk sequences. The general flow of the algorithm is as follows: $G = (V, E)$ is a given network and $f : v \rightarrow \mathbb{R}^d$ is the mapping function from nodes to feature representation. Here, $d$ is a hyperparameter representing the dimension of the vector and $f$ is a matrix of size $|V|/C_2^d$. For each source node $u \in V$, $N_s(u)$ is defined as a neighborhood of node $u$ generated through a neighborhood sampling strategy $S$. The problem translates to optimizing the following objective functions:

$$\max \sum_{i \in V} \log P_i(N_i(u) | f(u)).$$  \hspace{1cm} (Equation 8)

Two standard assumptions are made in order to make the optimization problem tractable: conditional independence:

$$P_i(N_i(u) | f(u)) = \prod_{n \in N_i(u)} P_i(n_i | f(u)),$$  \hspace{1cm} (Equation 9)

and symmetry in feature space:

$$P_i(n_i | f(u)) = \frac{\exp(f(n_i) \cdot f(u))}{\sum_{v \in V} \exp(f(v) \cdot f(u))}.$$  \hspace{1cm} (Equation 10)

| Relationship Type         | Database             | Number of Pairs |
|---------------------------|----------------------|-----------------|
| miRNA-lncRNA              | lncRNA2SNP2          | 8,374           |
| miRNA-disease             | HMDD                 | 16,427          |
| miRNA-protein             | miRTarBase           | 4,944           |
| IncRNA-disease            | lncRNA2Disease, lncRNA2SNP2 | 1,264         |
| IncRNA-protein            | lncRNA2Target        | 690             |
| Protein-disease           | DisGeNET             | 25,087          |
| Drug-protein              | DrugBank             | 11,107          |
| Drug-disease              | CTD                  | 18,416          |
| Protein-protein           | STRING               | 19,237          |
| Total                     | MAN                  | 105,546         |

| Node                      | Number of Nodes     |
|---------------------------|----------------------|
| Disease                   | 2,062                |
| IncRNA                    | 769                  |
| miRNA                     | 1,023                |
| Protein                   | 1,649                |
| Drug                      | 1,025                |
| Total                     | 6,528                |
Algorithm 1 The \textit{node2vec} Algorithm

\begin{algorithm}
\begin{algorithmic}
\State \textbf{LearnFeatures} (Graph $G = (V, E, W)$, Dimensions $d$, Walks per node $r$, Walk length $l$, Context size $k$, Return $p$, In-out $q$)
\State $\pi = \text{Preprocess Modified Weights} (G, p, q)$
\State $G' = (V, E, \pi)$
\State Initialize walks to Empty
\For {iter $= 1 \text{ to } r$}
\For {all nodes $u \in V$}
\State walk = \text{node2vecWalk} ($G'$, $u$, $l$)
\State Append walk to walks
\EndFor
\EndFor
\State $f = \text{Stochastic Gradient Descent} (k, d, \text{walks})$
\State return $f$
\State node2vecWalk (Graph $G' = (V, E, \pi)$, Start node $u$, Length $l$)
\State Initialize walk to $[u]$
\For {walk_iter $= 1 \text{ to } l$}
\State $curr = \text{walk} \left[ -1 \right]$
\State $V_{curr} = \text{Get Neighbors} (curr, G')$
\State $s = \text{Alias Sample} (V_{curr}, \pi)$
\State Append $s$ to walk
\EndFor
\State return walk
\end{algorithmic}
\end{algorithm}

With the above assumptions, the objective in Equation 8 simplifies to
\[
\max \sum_{x \in V} \left[ - \log Z_x + \sum_{n \in N(x)} f(n) \cdot f(u) \right]. \tag{Equation 11}
\]

Give the source node $u$ a random walk with length $l$, let $c_i$ be the $i$-th node of the walk, and the starting node $c_0 = u$. The node $c_i$ obeys the following distribution:
\[
P(c_i = x | c_{i-1} = v) = \begin{cases} 
\frac{\pi_{vx}}{Z} & \text{if } (v, x) \in E \\
0 & \text{otherwise}
\end{cases}, \tag{Equation 12}
\]
where $\pi_{vx}$ is the unnormalized transition probability between nodes $v$ and $x$, and $Z$ is the normalizing constant. Directly setting the transition probability to the edge weight $\pi_{vx} = \pi_{sv}$ cannot effectively consider the network structure and search for different neighbor spaces. The walk now needs to decide on the next step, so it evaluates the transition probabilities $\pi_{vx}$ on edges $(v, x)$ leading from $v$ by setting two parameters $p$ and $q$. Let $\pi_{vx} = \alpha_{pq}(t, x) \cdot w_{vx}$, where
\[
\alpha_{pq}(t, x) = \begin{cases} 
\frac{1}{p} & \text{if } d_{x} = 0 \\
1 & \text{if } d_{x} = 1 \\
\frac{1}{q} & \text{if } d_{x} = 2
\end{cases}, \tag{Equation 13}
\]
and $d_{x}$ is the shortest path distance between nodes $t$ and $x$.

The \textit{node2vec} algorithm is as follows:

\begin{algorithm}
\begin{algorithmic}
\State \textbf{LearnFeatures} (Graph $G = (V, E, W)$, Dimensions $d$, Walks per node $r$, Walk length $l$, Context size $k$, Return $p$, In-out $q$)
\State $\pi = \text{Preprocess Modified Weights} (G, p, q)$
\State $G' = (V, E, \pi)$
\State Initialize walks to Empty
\For {iter $= 1 \text{ to } r$}
\For {all nodes $u \in V$}
\State walk = \text{node2vecWalk} ($G'$, $u$, $l$)
\State Append walk to walks
\EndFor
\EndFor
\State $f = \text{Stochastic Gradient Descent} (k, d, \text{walks})$
\State return $f$
\State node2vecWalk (Graph $G' = (V, E, \pi)$, Start node $u$, Length $l$)
\State Initialize walk to $[u]$
\For {walk_iter $= 1 \text{ to } l$}
\State $curr = \text{walk} \left[ -1 \right]$
\State $V_{curr} = \text{Get Neighbors} (curr, G')$
\State $s = \text{Alias Sample} (V_{curr}, \pi)$
\State Append $s$ to walk
\EndFor
\State return walk
\end{algorithmic}
\end{algorithm}
Chen, X., Liu, C., Jiang, J., Li, W., Li, W., Liu, G., Zhou, W., Huang, J., and Tang, Y. (2012). Prediction of drug-target interactions and drug repositioning via network-based inference. PLoS Comput. Biol. 8, e1002503.

Chen, X. (2015). Predicting IncRNA-disease associations and constructing IncRNA functional similarity network based on the information of miRNA. Sci. Rep. 5, 13186.

Lin, C.-H., Konecki, D.M., Liu, M., Wilson, S.J., Nassar, H., Wilkins, A.D., Gleich, D.F., and Lichtarge, O. (2019). Multimodal network diffusion predicts future disease-gene-chemical associations. Bioinformatics 35, 1536–1543.

Barabási, A.-L., and Oltvai, Z.N. (2004). Network biology: understanding the cell’s functional organization. Nat. Rev. Genet. 5, 101–113.

Barabási, A.-L., and Albert, R. (1999). Emergence of scaling in random networks. Science 286, 509–512.

Strogatz, S.H. (2001). Exploring complex networks. Nature 410, 268–276.

Golub, G.H., and Reinsch, C. (1971). Singular value decomposition and least squares solutions. In Handbook for Automatic Computation, Linear Algebra, J.H. Wilkinson and C.H. Reinsch, eds. (Springer), pp. 134–151.

Roweis, S.T., and Saul, L.K. (2000). Nonlinear dimensionality reduction by locally linear embedding. Science 290, 2323–2326.

Perozzi, B., Al-Rfou, R., and Skiena, S. (2014). Deepwalk: Online learning of social representations. In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (ACM), pp. 701–710.

Grover, A., and Leskovec, J. (2016). node2vec: Scalable feature learning for networks. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (ACM), pp. 855–864.

Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., and Mei, Q. (2015). LINE: Large-scale network information embedding. In Proceedings of the 24th International Conference on World Wide Web (International World Wide Web Conference Committee), pp. 1067–1077.

Guo, Z.-H., Yi, H.-C., and You, Z.-H. (2019). Construction and Comprehensive Analysis of a Molecular Association Network via IncRNA-miRNA -Disease-Drug-Protein Graph. Cells 8, 866.

van Laarhoven, T., Nabuurs, S.B., and Marchiori, E. (2011). Gaussian interaction profile kernels for predicting drug-target interaction. Bioinformatics 27, 3036–3043.

Chen, X., Wang, C.-C., Yin, J., and You, Z.-H. (2018). Novel Human miRNA-Disease Association Inference Based on Random Forest. Mol. Ther. Nucleic Acids 13, 568–579.

Miao, Y.-R., Liu, W., Zhang, Q., and Guo, A.-Y. (2018). IncRNASNP2: an updated database of functional SNPs and mutations in human and mouse IncRNAs. Nucleic Acids Res. 46 (D1), D276–D280.

Huang, Z., Shi, J., Gao, Y., Cui, C., Zhang, S., Li, J., Zhou, Y., and Cui, Q. (2019). HMDD v3.0: a database for experimentally supported human microRNA-disease associations. Nucleic Acids Res. 47 (D1), D1013–D1017.

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., et al. (2018). miR TarBase update 2018: a resource for experimentally validated microRNA-target interactions. Nucleic Acids Res. 46 (D1), D296–D302.

Chen, G., Wang, Z., Wang, D., Qiu, C., Liu, M., Chen, X., Zhang, Q., Yan, G., and Cui, Q. (2013). LncRNA Disease: a database for long non-coding RNA-associated diseases. Nucleic Acids Res. 41 (D1), D983–D986.

Cheng, L., Wang, P., Tian, R., Wang, S., Guo, Q., Luo, M., Zhou, W., Liu, G., Jiang, H., and Jiang, Q. (2019). LncRNA2Target v2.0: a comprehensive database for target genes of IncRNAs in human and mouse. Nucleic Acids Res. 47 (D1), D140–D144.

Piñero, J., Bravo, À., Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., García-Garcia, J., Sanz, F., and Furlong, L.I. (2017). DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res. 45 (D1), D833–D839.

Wishart, D.S., Feunang, Y.D., Guo, A.C., Lo, E.J., Marcu, A., Grant, J.R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., et al. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 46 (D1), D1074–D1082.

Davis, A.P., Grondin, C.I., Johnson, R.I., Sciaky, D., McMorrnan, R., Wiegers, J., Wiegers, T.C., and Mattingly, C.J. (2019). The comparative toxicogenomics database: Update 2019. Nucleic Acids Res. 47 (D1), D948–D954.

Szklarczyk, D., Morris, J.H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N.T., Roth, A., Bork, P., et al. (2017). The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. Nucleic Acids Res. 45 (D1), D362–D368.

Kozomara, A., Birgaoanu, M., and Griffiths-Jones, S. (2019). miRBase: from microRNA sequences to function. Nucleic Acids Res. 47 (D1), D155–D162.

Fang, S., Zhang, L., Guo, J., Niui, Y., Wu, Y., Li, H., Zhao, L., Li, X., Teng, X., Sun, X., et al. (2018). NONCODEv5: a comprehensive annotation database for long non-coding RNAs. Nucleic Acids Res. 46 (D1), D308–D314.

Shen, J., Zhang, J., Luo, X., Zhu, W., Yu, K., Chen, K., Li, Y., and Jiang, H. (2007). Predicting protein-protein interactions based only on sequences information. Proc. Natl. Acad. Sci. USA 104, 4337–4341.

Wang, D., Wang, J., Li, M., Song, F., and Cui, Q. (2010). Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. Bioinformatics 26, 1644–1650.

Landrum, G. (2013). Rdkit documentation. Release 2013, 1–79, https://buildmedia.readthedocs.org/media/pdf/rdkit/latest/rdkit.pdf.

Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J. Chem. Inf. Comput. Sci. 28, 31–36.