HNUE JOURNAL OF SCIENCE DOI: 10.18173/2354-1059.2023-0026
Natural Sciences 2023, Volume 68, Issue 2, pp. 121-130
This paper is available online at http://stdb.hnue.edu.vn

PREDICTING MICRORNA-DISEASE ASSOCIATIONS USING HETEROGENEOUS GRAPH REPRESENTATION LEARNING

Le Thi Tu Kien and Le Xuan Hien
Faculty of Information Technology, Hanoi National University of Education

Abstract. MicroRNAs (miRNAs) are small non-coding RNAs that play a crucial role in regulating gene expression post-transcriptionally. These molecules have been implicated in various diseases, including cancer, viral infections, cardiovascular disorders, and neurodegenerative conditions. This study introduces a novel approach for predicting miRNA-disease associations by leveraging heterogeneous graph representation learning. By integrating both structural and semantic information from the heterogeneous graph, our method offers an enhanced prediction process for discerning the relationship between miRNAs and diseases. Our experimental findings demonstrate the effectiveness of our prediction method, yielding promising results with an average AUC value of 0.907.

Keywords: miRNA, disease, miRNA-disease relationship, graph representation learning, heterogeneous graph.

1. Introduction

MicroRNA (miRNA) is a non-coding RNA with a length of about 22 nucleotides, which often has the function of inhibiting the expression of some genes. Recently, many studies identified miRNA as one of the important components in cells and play a key role in many different basic biological processes. Therefore, changes in miRNA function are related to many different types of diseases. Searching for the relationship between miRNAs and diseases on a large scale has become an important goal in biomedical research. This promotes understanding of diseases at the molecular level and brings benefits to human disease treatment and prevention. However, the current understanding of the association between miRNAs and diseases is limited. Experimental identification of diseases related to miRNA through existing biological techniques is costly and time-consuming. However, with a large amount of biological data on miRNAs being generated, researchers can build powerful computational methods that can detect potential associations between miRNAs and diseases [1, 2].

Some computational methods to predict miRNAs related to disease have been proposed. Lu et al. [3] analyzed data on the relationship between miRNAs and diseases

Received June 5, 2023. Revised June 22, 2023. Accepted June 30, 2023.
Contact Le Thi Tu Kien, e-mail address: kienltt@hnue.edu.vn

and proposed several patterns of relationship between miRNAs and human diseases. This has set a new foundation for research on miRNAs related to disease and has provided support for research on diseases at the miRNA level. Based on the assumption that diseases similar in phenotype tend to have relationships with functionally related miRNAs, Zhang and colleagues built the first method to predict the set of diseases related to miRNAs. This method identifies potential miRNAs related to cardiovascular disease by integrating information from known miRNA sets and Gene Ontology. The fact that this method relies heavily on the miRNA set so limits its applicability. Jiang et al. [4] built a computational method based on hypergeometric distribution to identify miRNAs related to diseases by integrating functional miRNA interaction network, disease similarity network, and known miRNA network, which includes a set of associations between miRNAs and diseases verified by experiments taken from the miR2Disease database. However, the functional miRNA network built only has information about the nearest neighbors of each miRNA used in calculating the relationship weight. Taking full advantage of similarity information in the global network will improve the accuracy of this algorithm. Liu et al. [5] proposed a PBMDA prediction model that integrates the known associations between miRNAs and human diseases, the functional similarity of miRNAs, the semantic similarity of diseases, and the Gaussian interaction profile kernel for miRNAs and diseases. They built a heterogeneous graph and continued to apply the depth-first search algorithm to find possible associations between miRNAs and diseases. Chen et al. [6] presented a GIMDA model to predict the associations between miRNAs and diseases by measuring the graphlet interaction between miRNAs and between diseases. Graphlet is a type of subgraph with few connections in a large network. GIMDA achieves decision performance but is very time-consuming.

In recent years, a new approach has been used to encode the structural information of graphs into vector features, which is graph representation learning or graph embedding [7-16]. The idea of this approach is to find a mapping to encode the information of a vertex (or a subgraph) into a point in a low-dimensional vector space, such that it optimizes the preservation of structural information between vertices in the old and new spaces. Then, the data in the form of vector features will be used to train machine learning models, to solve classification and prediction problems on graph data. The difference of this approach is that it considers the representation of graph data as a machine learning problem to optimize the preservation of structural information, while previous methods only consider it as a data preprocessing step. The current graph representation learning methods are mainly developed for problems in the field of social network mining and scientific paper citation network. Therefore, in this study, we aim to investigate graph representation learning methods and apply them to solve the problem of predicting links between miRNAs and diseases.

2. Content

2.1. Propose a framework for predicting the relationship between miRNAs and diseases based on heterogeneous graph representation learning

In this study, we build a framework to predict the associations between miRNAs and diseases based on the heterogeneous graph representation learning method (Figure 1). In

the first step, we construct a heterogeneous graph between miRNAs and diseases from databases on the associations between miRNAs and diseases, functional similarity between miRNAs, and semantic similarity of diseases (Figure 1a). Thus, the network constructed includes 2 types of nodes and 3 types of edges. In step two, we find meta-paths between nodes in the heterogeneous graph constructed in step 1 (Figure 1b). Then, from the set of meta-paths obtained, we construct random walks through the vertices of the graph combined with using the heterogeneous skip-gram model to find feature vectors for each vertex of the graph (Figure 1c). In step four, we construct an initial value matrix **P** between miRNAs and diseases from the feature vectors of the nodes found in step 3 (Figure 1d). In the next step, from the miRNAs and diseases association network, we construct a label matrix **L** with label 1 as having a relationship and label 0 as having no relationship between miRNAs and diseases (Figure 1e). Combine matrix **P** and **L** to split data into two training and testing sets. From the training data set, we build a logistic gegression prediction model C (Figure 1f). Finally, apply the trained model C in step 6 to predict the relationship between miRNAs and diseases. In the following sections, we describe in detail how to construct a heterogeneous graph between miRNAs and diseases, the technique to find meta-paths, and the method to learn heterogeneous graph representation.

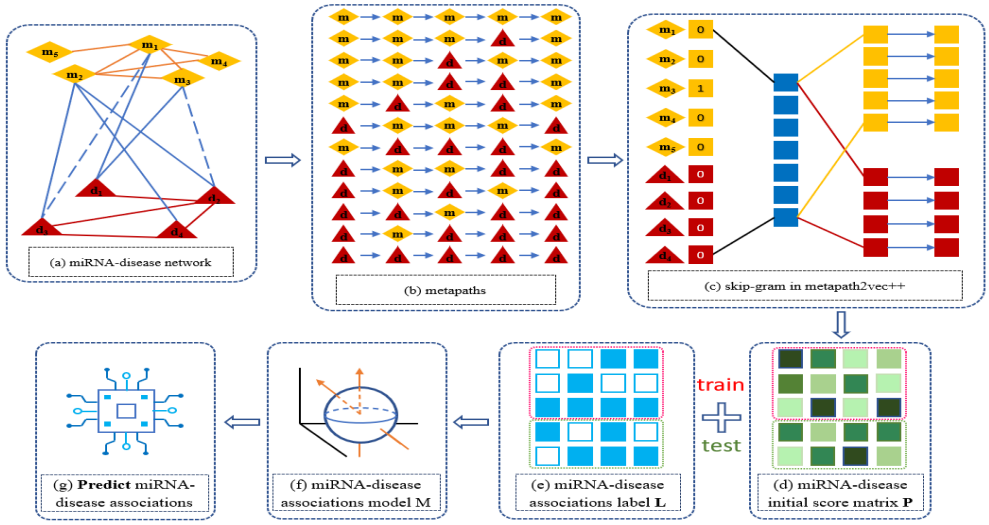

Figure 1. Propose a framework for predicting the relationship between miRNAs and diseases based on heterogeneous graph representation learning

2.1.1. Construction of heterogeneous graph between miRNAs and disease

A heterogeneous graph is defined as $G = (V, E, T)$, where each vertex $v \in V$ and each edge $e \in E$ are associated with corresponding mapping functions of $\psi(v): V \rightarrow T_V$ and $\varphi(e): E \rightarrow T_E$. T_V and T_E are respectively the set of vertex types and the set of edge types satisfying the condition $|T_V| + |T_E| > 2$ [8]. The heterogeneous network between miRNAs and diseases in our problem consists of two types of vertices miRNAs and diseases, three types of edges representing the relationship between miRNAs and diseases, the relationship between a pair of miRNAs, and the relationship between a pair of diseases. The way to construct these three types of relationships is described below.

Firstly, from the data on the association between miRNAs and diseases, we construct a network. This network consists of a set of vertices $V_M = \{m_1, m_2, ..., m_n\}$ representing a set of n miRNAs and a set of vertices $V_D = \{d_1, d_2, ..., d_k\}$ representing a set of k diseases. The miRNA-disease relationship network is stored in the adjacency matrix L. If miRNA m_i is linked to disease d_j with weight c, with $0 < c \leq 1$, then $\boldsymbol{L}(i,j) = 1$, otherwise $\boldsymbol{L}(i,j) = 0$. In Figure 1a, the link between miRNAs and diseases is represented by solid blue lines.

Secondly, the MFSN (MiRNA Functional Similarity Network) represents the functional similarity between miRNAs through the weight values describing the functional similarity between them. This is a homogeneous network consisting of a set of vertices $V_M = \{m_1, m_2, ..., m_n\}$ representing n miRNAs. The weight for each pair of miRNAs is calculated based on the observation that genes with functional similarity often have relationships with similar diseases. Two vertices m_i and m_j are connected by an edge in the network if the ratio of functional similarity between miRNA i and j is greater than a certain threshold. In this research, the value of the threshold is greater than 0. The weight of functional similarity is also used as the weight for the edge on the MFSN graph. The MFSN network is stored in matrix **M** with the value at row *i* and column *j* indicating the ratio of functional similarity between miRNA *i* and *j*. In Figure 1a, each miRNA is represented by a yellow vertex, yellow edges connecting these vertices represent the similarity between two miRNAs

Thirdly, the DSSN (Disease Semantic Similarity Network) represents the semantic similarity weights between diseases. The network has a set of vertices $V_D = \{d_1, d_2, ..., d_k\}$ representing a set of k diseases. Two vertices d_i and d_j are connected by an edge in the network if the ratio of semantic similarity between them is greater than a certain threshold. In this research, the value of the threshold is greater than 0. We use the Disease Ontology database [17] and DOSE (Disease Ontology Semantic and Enrichment) library [18] to reference Disease Ontology ID (DOID) and to calculate the semantic similarity score between two diseases. The DSSN network is stored in matrix **D** where element $D(i,j)$ of the matrix at row *i* and column *j* indicates the ratio of semantic similarity between two diseases *i* and *j*.

2.1.2. Learning feature vectors for nodes in the miRNAs and diseases heterogeneous graph

Heterogeneous network representation learning is a method to find a representation X in a lower-dimensional latent space, with $X \in R^{|V| \times d}$ and $d \ll |V|$, such that X contains the structure and semantic information of the graph G. In this case, X is a matrix with the number of rows equal to the total number of vertices in G, and each row v^{th} corresponds to a *d*-dimensional vector X_v.

In 2017, Yuxiao Dong et al. proposed a heterogeneous graph representation learning method based on two techniques: meta-path-based random walks and heterogeneous skip-gram [8]. The meta-path-based random walk technique is used to integrate semantic and structural information between vertices in heterogeneous graphs into the process of learning vertex representations as vectors. Suppose, ρ is a meta-path on the heterogeneous graph G defined as follows:

$$V_1 \xrightarrow{R_1} V_2 \xrightarrow{R_2} \cdots V_t \xrightarrow{R_t} V_{t+1} \xrightarrow{R_{t+1}} \cdots V_l$$

where R_i is represent the relationship between two types of nodes in V. At step i, the transition probability $tp(v^{i+1}|v_t^i, \rho)$ represents the probability for node i of type t to move to the next point i + 1 on the meta-path ρ. This transition probability is calculated by the following formula:

$$tp(v^{i+1}|v_t^i, \rho) = \begin{cases} \frac{1}{|N_{t+1}(v_t^i)|}, & (v^{t+1}, v^i) \in E; \varphi(v^{i+1}) = \mathrm{t}+1; \\ 0, & (v^{t+1}, v^i) \in E; \varphi(v^{i+1}) \neq \mathrm{t}+1; \\ 0, & (v^{t+1}, v^i) \notin E; \end{cases} \quad (1)$$

where $v_t^i \in V_t$, $N_{t+1}(v_t^i)$ denotes the (t+1) type of neighborhood of the node v_t^i and $\varphi(v^{i+1})$ is the type of node v^{i+1}.

Heterogeneous skip-gram is a technique for learning node representations in a heterogeneous graph G. The goal of the technique is to find the maximum probability in the heterogeneous context $N_t(v)$ of the node *v*:

$$\arg max \max_{\theta} \sum_{v \in V} \sum_{t \in T_V} \sum_{b_t \in N_t(v)} \log p(b_t|v;\theta) \quad (2)$$

where $t \in T_V$ and $N_t(v)$ are the neighborhood of *v* with node type *t*. The component $p(b_t|v;\theta)$ is a softmax function adjusted according to each node type:

$$p(b_t|v;\theta) = \frac{e^{X_{b_t} X_v^T}}{\sum_{u_t \in V_t} e^{X_{u_t} X_v^T}}, \quad (3)$$

where V_t denotes a set of nodes of type t, X is the embedded matrix, and X_v, X_{b_t}, X_{u_t} denote the rows *v*, b_t and u_t in the matrix X.

Because the number of nodes in graph G is usually large, the computational cost of E.q 3 will also become very large. To solve this problem, we use the negative sampling method to reduce the computational size. If node b_t is a local neighbor of node *v*, events are considered a mixture of two independent events. The first event isb_t, and *v* appears in the list at the same time. The second event is the set Z of noise nodes does not appear in the list at the same time as node *v*. Therefore, the objective function of the heterogeneous graph representation learning method is:

$$O(X) = \log \sigma(X_{b_t} X_v^T) + \sum_{z=1}^{Z} E_{u_t^z \sim P_t(u_t)} [\log \sigma(-X_{u_t^z} X_v^T)], \quad (4)$$

where $\sigma(x) = 1/(1 + e^{-x})$ is the sigmoid function. $P_t(u_t)$ is the negative distribution probability. $E_{u_t^z \sim P_t(u_t)}$ is an expectation when u_t^z follows the distribution $P_t(u_t)$, with u_t^z is a node of type t of the set of nodes obtained from the negative method. The initial computational scale is V but is now reduced to Z (Z ≪ V).

In Figure 1, the heterogeneous graph between miRNAs and diseases, we apply the meta-path based random walk technique to find meta-paths (Figure 1b). Then, these meta-paths are combined into the heterogeneous skip-gram learning model to obtain a set of feature vectors of the vertices in the graph (Figure 1c).

2.2. Experiment results

2.2.1. Data

We obtained data on the association between miRNAs and diseases from the HMDD database [19]. The miRNAs produce the same mature miRNA were grouped together.

For example, the miRNAs hsa-let-7a-1, hsa-let-7a-2, hsa-let-7a-3 were grouped into the hsa-let-7a cluster. The mature miRNAs then map into a single miRNA gene. The disease names are also standardized using the MeSH database terms. Finally, the dataset includes 495 miRNAs, 383 diseases and 5430 relationships. Based on this dataset, we build networks of relationships between miRNA-disease (MDAN), functional similarity network of miRNAs (MFSN), and semantic similarity network of diseases (DSSN). These three networks merge into a heterogeneous network as discussed in Section 2.1.1.

2.2.2. Meta-paths and feature vectors

The meta-paths in the heterogeneous graph need to ensure that the endpoints must be of the same type of vertex and usually one meta-path is used for each run [8]. In the experiment, we applied techniques as mentioned in section 2.1.2 to get meta-paths and feature vectors. The meta-path technique based on a random walk is used to obtain 12 meta-paths from the heterogeneous graph G. These meta-paths met the criterion mentioned above. Next, we use the meta-paths as the input of the heterogeneous skip-gram model. The initialization matrix X with 878 rows is equal to the number of vertices of the graph. We run the algorithm with one meta-path or a set of meta-paths and obtain feature vectors for 878 nodes.

2.2.3. Training prediction model

We build the initialization value matrix **P** based on feature vectors in the **X** (Figure 1d). $\boldsymbol{P} = X_{disease}X_{miRNA}^{T}X_{miRNA}$, where $X_{disease}$ and X_{miRNA} are the rows of matrix **X**. The label matrix **L** presents relationships between miRNAs and diseases (Figure 1e) as discussed in section 2.1.1. We divide two matrices **P** and **L** into training and testing datasets with corresponding ratios of α and $1 - \alpha$. We train a classification model M by the logistic regression learning method (Figure 1f) and then use it to repredict associations between miRNAs and diseases for the test set (Figure 1g). We compare the predicted values with the actual label values of the testing data to evaluate the accuracy of the model. Apply the LOOCV (leave-one-out cross validation) evaluation method for 5430 known and experimentally validated relationships between miRNAs and diseases. We used the AUC metric and ROC curve to evaluate the results.

2.2.4. Results

The experiment was divided into three parts corresponding to the input factors of the graph representation learning model that affect the results of the problem of predicting the relationship between miRNAs and diseases.

Experiment 1 - The effect of the meta-path structure on the prediction results

With the input parameter value of 12 meta-paths, we performed the steps in the experimental procedure described in Section 2.2.3. Finally, we performed cross-validation and obtained the result as the ROC curve along with the AUC value as shown in Figure 2. We found that the algorithm depends heavily on the input metapaths. Table 1 describes the AUC value results in the ROC curves of metapaths m_1, m_2, m_3, m_4. With meta-paths m_1, m_2, m_3, we repeated from 2 to 3 times of the substructure meta-paths then the AUC value of the model was only in the range of 0.7 to 0.8. But with meta-path m_4 which is a combination of m_1, m_2, m_3, the result returned with an AUC value above 0.9.

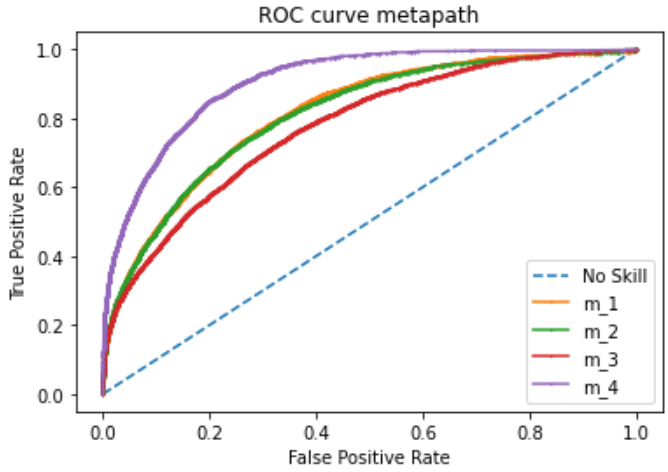

Figure 2. ROC curves present the accuracy of the classifiers with different meta-paths

Table 1. The AUC and running time of meta-paths m_1, m_2, m_3, m_4

Metapath	m_1	m_2	m_3	m_4
AUC	0.794	0.812	0.777	0.907
Feature vector finding time (s)	184.9	250.9	264.29	260.4
Prediction time(s)	7.88	7.13	7.92	7.71

Experiment 2 - The effect of the meta-path length on the prediction results

From the experiment 1 results, we find that the meta-path m_4 has the best AUC. Therefore, in experiment 2, we based on meta-path m_4 to evaluate the influence of the meta-path length. We set the value of walk_length parameter to 32, 64, 128, 256. The experimental results in Table 2 and Figure 3 show that the accuracy of the classifier is little affected when changing the walk_length parameter. This also reflects that the biological graph we are considering has low complexity, not diverse in terms of quantity and type of vertices and edges. But the running time is proportional to the length of the meta-path.

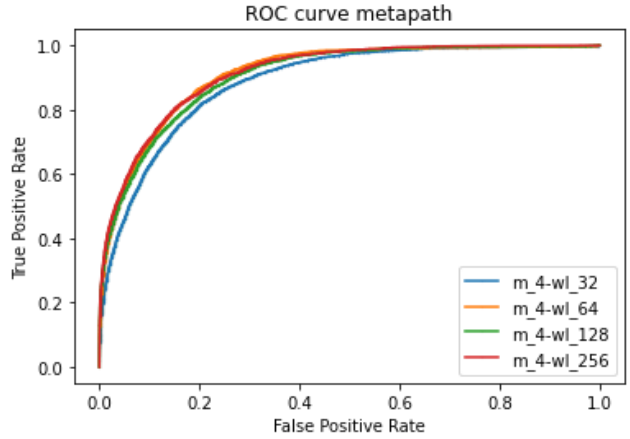

Figure 3. ROC curves present the accuracy of the classifiers with different meta-path lengths

Table 2. The AUC and running time with different meta-path lengths

Metapath	m_4-wl_32	m_4-wl_64	m_4-wl_128	m_4-wl_256
AUC	0.874	0.913	0.906	0.907
Feature vector finding time (s)	87.9	164.1	261.8	443.9
Prediction time(s)	8.04	8.04	8.08	8.5

Experiment 3 - The effect of the feature vector dimension on the prediction results

In experiment 3, except for the parameter of the feature vector dimension, we keep the input parameters of the problem constant and used the meta-path m_4. The parameter dim, the dimension of the feature vector, is set to values 20, 50, 100, and 200. The results in Table 3 and Figure 4 show that the larger the dimension of the graph feature vector, the more accurate the classifier is. In addition, the running time increases slowly.

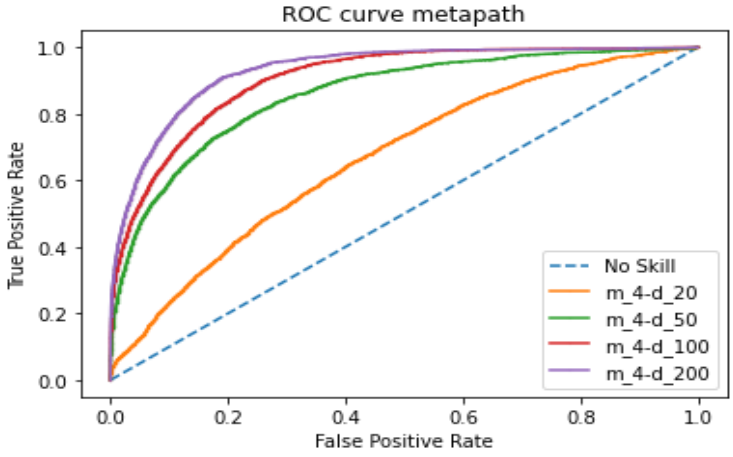

Figure 4. ROC curves present the accuracy of the classifiers with different feature vector dimensions

Table 3. The AUC and running time with different feature vector dimensions

Metapath	m_4–d_20	m_4–d_50	m_4–d_100	m_4–d_200
AUC	0.669	0.859	0.905	0.931
Feature vector finding time (s)	231.8	230.6	232.6	251.9
Prediction time(s)	5.12	5.88	7.78	13.38

In summary, the meta-path structure has a great effect on the prediction results, but its path length has no effect. It recommends that the input graph should have more additional links and the type of links between nodes. Secondly, the number of dimensions of the feature vectors has an effect on the results of the prediction associations and it is also related to the running time of the algorithm. Therefore, in the future, we need to study to enrich the association graph between miRNAs and diseases. Then, we setup reasonable parameter values to build a model to predict better the association between miRNAs and diseases.

3. Conclusions

In this study, we propose a framework for predicting associations between miRNAs and diseases based on the heterogeneous graph representation learning model. The experimental results show that our method gives good prediction performance with an average AUC value of 0.907. The experiments only evaluate the influence of some factors such as structure, length of meta-paths, and dimension of feature vectors on the accuracy of the link prediction problem. Therefore, in the future research, we will analyze and evaluate the new predicted relationships between miRNAs and diseases. In addition, we need to integrate more related biological data such as protein, cirARN, … to enrich the information for the graph and bring more accurate prediction results.

***Acknowledgments*.** This work is supported by the Vietnam Ministry of Education and Training, project No. B2021-SPH-01.

REFERENCES

[1] Nguyen VT, Le TTK, Nguyen TQV, Tran DH, 2021. Inferring miRNA-disease associations using collaborative filtering and resource allocation on a tripartite graph. *BMC Medical Genomics*, https://doi.org/10.1186/s12920-021-01078-8.

[2] N. D. Hung, T. T. Tien and T. D. Hung, 2015. Prediction of the association between miRNAs and diseases by RWRS. *Journal of Science of HNUE*, Vol. 60, pp. 10-20 (in Vietnamese).

[3] M. Lu, Q. Zhang, M. Deng, J. Miao, Y. Guo, W. Gao and Q. Cui, 2008. An Analysis of Human MicroRNA and Disease Associations. *PloS One* 3(10): e3420. doi: 10.1371/journal.pone.0003420. Epub 2008 Oct 15. PMID: 18923704; PMCID: PMC2559869.

[4] Q. Jiang, Y. Hao, G. Wang, L. Juan, T. Zhang, M. Teng, Y. Liu and Y. Wang, 2010. Prioritization of disease microRNAs through a human phenome-microRNAome network. *BMC Systems Biology*, Vol. 4, p. s2.

[5] Liu Y, Zeng X, He Z, Zou Q, 2017. Inferring MicroRNA-disease associations by random walk on a heterogeneous network with multiple data sources. *IEEE/ACM Trans Comput Biol Bioinform*, Vol. 14, No. 4, pp. 905-915.

[6] X. Chen, NN. Guan , JQ. Li, GY. Yan, 2018. GIMDA: graphlet interaction based MiRNA-disease association prediction. *J. Cell. Mol. Med.*, Vol. 22, No. 3, pp. 1548-1561.

[7] A. Bordes, N. Usunier and A. Garcia-Duran, 2013. Translating Embeddings for Modeling Multi-relational Data. *Proceeding of Advances in Neural Information Processing Systems 26 (NIPS 2013)*. ISBN: 9781632660244.

[8] Y. Dong, N. V. Chawla and A. Swami, 2017. Metapath2vec: Scalable Representation Learning for Heterogeneous Networks. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 135-144.

[9] Y. Goldberg and O. Levy, 2014. Word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. CoRR abs/1402.3722.

[10] A. Grover and J. Leskovec, 2016. Node2vec: Scalable feature learning for networks. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 855-864.

[11] J. Tang, M. Qu and Q. Mei, 2015. PTE: Predictive Text Embedding through Large-scale Heterogeneous Text Networks. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1165-1174.

[12] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan and Q. Mei, 2015. LINE: Large-scale Information Network Embedding. Proceedings of the 24th International Conference on World Wide Web, pp. 1067-1077.

[13] W. L. Hamilton, R. Ying and J. Leskovec, 2017a. Representation learning on graphs: Methods and applications. *IEEE Data Eng. Bull.*

[14] P. D. Hoff, A. E. Raftery and M. S. Handcock, 2006. Latent space approaches to social network analysis. *JASA*, Vol. 43, No. 4, pp. 439-561.

[15] Imran M, Yin H, Chen T, Huang Z and Zheng K, 2022. DeHIN: A Decentralized Framework for Embedding Large-Scale Heterogeneous Information Networks. *IEEE Transactions on Knowledge and Data Engineering.* 10.1109/TKDE.2022.3141951. 35:4.(3645-3657).

[16] Roy A, Mittal S and Chakraborty T., 2022. MG2Vec+: A multi-headed graph attention network for multigraph embedding. *Knowledge and Information Systems.* 10.1007/s10115-022-01706-4. 65:1. (111-132). Online publication date: 1-Jan-2023.

[17] W. A. Kibbe, C. Arze, V. Felix, E. Mitraka, E. Bolton, G. Fu, C. J. Mungall, J. X. Binder, J. Malone, D. Vasant, H. Parkinson and L. M. Schriml, 2015. Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Research*, Vol. 43, pp. 1071-1078.

[18] G. Yu, L.G. Wang, G.-R. Yan and Q.Y. He, 2015. *DOSE:* an R/Bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics*, Vol. 31, pp. 608-609.

[19] Huang Z, Shi J, Gao Y, Cui C, Zhang S, Li J, Zhou Y, Cui Q, 2019. *HMDD v3.0: a database for experimentally supported human microRNA-disease associations*. *Nucleic Acids Res*, 47(D1):D1013-D1017. doi: 10.1093/nar/gky1010. PMID: 30364956; PMCID: PMC6323994.