LINK PREDICTION IN MICRORNA-MEDIATED BIOMOLECULAR NETWORKS

HUANG YUAN

PhD

The Hong Kong Polytechnic University

2020

The Hong Kong Polytechnic University

Department of Computing

LINK PREDICTION IN MICRORNA-MEDIATED BIOMOLECULAR NETWORKS

Huang Yuan

A thesis submitted in partial fulfilment of

the requirements for the degree of

Doctor of Philosophy

May 2020

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgment has been made in the text.

_____

HUANG YUAN

ABSTRACT

Many problems in the real-world can be formulated as discovering the existence of relationship between objects in a set of inter-related objects. For example, in molecular biology, it is known that microRNA and human diseases are related as they may interact with each other. While the existence of interaction relationship between some of them may be known, the existence of some others may not. One problem is, therefore, for the existence of interaction relationship between a microRNA and a human disease to be determined based on known relationship between other microRNAs and human diseases. If we represent microRNAs and human diseases as nodes in a network, then the links between them can be used to represent their interaction relationship, we have a biomolecular network. Given such a network, we can then define a link prediction problem as the prediction of missing links in the network based on existing links.

In this thesis, we tackle the link prediction problem of three kinds of biomolecular networks that involve mediated microRNA. Specifically, we predict three types of interaction relationships between microRNA and three other different objects: (i) complex human diseases, (ii) drug resistance and (iii) lncRNA. Based on known interaction data obtained from public databases, we construct microRNA-mediated biomolecular networks containing nodes and unweighted links. The nodes are of two types. One type represents microRNA and the other represents either diseases, drug resistance or lncRNA. The links between these different types of nodes represent

interaction relationship between the two types of objects. Given the biomolecular networks, our problems are to use the known links to predict the missing ones in the networks.

In the datasets we collected, known interaction data are often limited in number. To improve the prediction performance, in addition to the known links, we introduce node information data that are biologically relevant to the objects that the nodes represent for link prediction. These data can be related to the biological or physicochemical properties of the objects. They can be concerned with expression profiles, drug structural data, RNA sequences, etc., and their data types can be very different. For example, when predicting links in microRNA-disease association network, the data we use to characterize the node of microRNAs can be another network -- the lncRNA-microRNA interaction network. When predicting the links in microRNA-drug resistance association network, the data we use to characterize the nodes of drugs and microRNAs are high-dimensional numerical features. When predicting the links in microRNA-lncRNA interaction network, the data we use to characterize the nodes of microRNA and lncRNA are network multiple similarity matrixes. The main challenges of our research, therefore, lie in finding ways to introduce these different kinds of node information during the prediction process. To overcome these difficulties, we propose four different algorithms that can each effectively tackle different challenges.

Specifically, to predict associations between microRNA and diseases, MVMTMDA algorithm considers the data incompleteness of lncRNA-microRNA interactions. It

formulates the prediction task as a multi-task problem, in which the links of lncRNA-microRNA interaction and microRNA-disease association are simultaneously predicted, and adopts multi-view learning to learn the embedding of microRNA nodes from two networks. When predicting the associations between microRNA and drug resistance, the nodes have attributes whose dimensions are up to thousands, which is extremely high. GCMDR algorithm used a spectral graph convolution technique to solve this problem. The deep neural network structure it adopts can be applied to high dimensional node numerical features, allowing an end-to-end prediction without any data preprocessing process. Different from other prediction tools for microRNA targets that are based on sequence matching, EPLMI algorithm for the first time, reformulates the lncRNA-microRNA interaction prediction task as a link prediction problem and adopts a two-way diffusion method to perform prediction. To improve the prediction performance of EPLMI, we further propose LMNLMI algorithm which use a similarity network fusion technique to collectively consider multiple types of lncRNA/microRNA similarity. The proposed algorithms have been applied on real-world datasets that we collected from the public databases. The experimental results illustrate our proposed models are accurate, efficient, robust to parameter settings and outperform state-of-the-art approaches.

PUBLICATIONS ARISING FROM THE THESIS

1. **Huang Y-A**, Hu P, Chan KC, You Z-H: Graph convolution for predicting associations between miRNA and drug resistance. ***Bioinformatics*** 2020, 36(3):851-858.

2. **Huang Y-A**, Chan KC, You Z-H: Constructing prediction models from expression profiles for large scale lncRNA–miRNA interaction profiling. ***Bioinformatics*** 2018, 34(5):812-819.

3. **Huang Y-A**, Chan KC, You Z-H, Hu P: Predicting microRNA-disease associations from lncRNA-microRNA interactions via multi-view multi-task learning. ***Briefings in Bioinformatics*** 2020.

4. Hu P, **Huang Y-A**, Chan KC, You Z-H: Learning Multimodal Networks from Heterogeneous Data for Prediction of lncRNA-miRNA Interactions. ***IEEE/ACM Transactions on Computational Biology and Bioinformatics*** 2019.

5. Chen X, **Huang Y-A**, You Z-H, Yan G-Y, Wang X-S: A novel approach based on KATZ measure to predict associations of human microbiota with non-infectious diseases. ***Bioinformatics*** 2017, 33(5):733-739.

6. Wang L, You Z-H, **Huang Y-A**, Huang D-S, Chan KC: An efficient approach based on multi-sources information to predict circRNA–disease associations using deep convolutional neural network. ***Bioinformatics*** 2020, 36(13):4038-4046.

7. Huang Z-A, **Huang Y-A**, You Z-H, Zhu Z, Sun Y: Novel link prediction for large-scale miRNA-lncRNA interaction network in a bipartite graph. ***BMC medical genomics*** 2018, 11(6):17-27.

8. Sun Y, Zhu Z, You Z-H, Zeng Z, Huang Z-A, **Huang Y-A**: FMSM: a novel computational model for predicting potential miRNA biomarkers for various human diseases. ***BMC systems biology*** 2018, 12(9):121.

9. Hu P, **Huang Y-A**, Chan KC, You Z-H: Discovering an Integrated Network in Heterogeneous Data for Predicting lncRNA-miRNA Interactions. In: ***International Conference on Intelligent Computing***: 2018. Springer: 539-545.

10. Wang L, You Z-H, Li Y-M, Zheng K, **Huang Y-A**: GCNCDA: A new method for predicting circRNA-disease associations based on Graph Convolutional Network Algorithm. ***PLOS Computational Biology*** 2020, 16(5):e1007568.

11. Zheng K, You Z-H, Li J-Q, Wang L, Guo Z-H, **Huang Y-A**: iCDA-CGR: Identification of circRNA-disease associations based on Chaos Game Representation. ***PLOS Computational Biology*** 2020, 16(5):e1007872.

12. You Z-H, Huang W-Z, Zhang S, **Huang Y-A**, Yu C-Q, Li L-P: An efficient ensemble learning approach for predicting protein-protein interactions by integrating

protein primary sequence and evolutionary information. ***IEEE/ACM transactions on computational biology and bioinformatics*** 2018, 16(3):809-817.

ACKNOWLEDGEMENTS

Firstly, I would like to give my sincere gratitude to Professor Keith C.C. Chan, my supervisor who, with extraordinary patience and consistent encouragement, provides advice of great value, inspiration of new ideas, and much help in researching. Without his strong support, this thesis would not been the present form.

Then, I would express my thankfulness to Prof. Zhu-Hong You, who always gives much support to me in academic and life. Besides, I am extremely grateful for Dr. Yan Liu's help.

I would like to thank all my friends and colleagues: Prof. Henry Leung, Dr. Pengwei Hu, Dr. Tiantian He, Dr. Zimu Zheng, Dr. Jiaxing Shen, Dr. Yanwen Wang, Dr. Bo Tan, Mr. Yuxin Bo, Mr. Zhe Li.

Special thanks to my parents and brother for all the supports and scarifies that you have made for me. Your love to me sustained thus far.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

Relational data among objects can be formulated as networks containing nodes and links in between. Predicting new links in the known networks is a fundamental task for applications in various domains. With the advent of big data storage solutions, data of different kinds are much easier to be obtained than before, which helps provide more information for describing the research objects and thus the opportunity to improve the prediction performance, however, but also brings new challenges that require new algorithms to be developed. Nowadays, the real relational data are often large-scale, high-dimension, incomplete and multisource, and thus the conventional link prediction algorithms like PageRank are hard to handle [1].

In the field of bioinformatics, knowledge from experimental and clinical discovering is often a network representation of relationships amongst a group of biomolecules, in which nodes represent the universal biological entities and links depict their interrelationship. Biomolecules make up cell signaling pathways, which interact with one another to form networks in their functional mechanism. As a cornerstone of systems biology, modelling signaling networks requires a combination of experimental and theoretical approaches including the development and analysis of models. In the recent years, along with the research upsurge in the field of noncoding RNA, microRNA has been intensively studied as 'star molecule' [2]. It is a cluster of ~22nt ncRNAs, which generally bind to the 3'UTR of the mRNA imperfectly [3]. In most cases, this can lead to translational inhibition or degradation of its target mRNA. Given the

ubiquitous regulation of microRNA on genes, it is important to decipher their function mechanism and the key is to detect the biomolecular networks that they involve [4]. The study of microRNA is still on its initial stage, and thus the microRNA-mediated biomolecular network is sparse [5]. It is an urgent need to predict the new links inside the networks and many attempts have been made to do so. Different types of methods and computational tools have been proposed to predict the genes and diseases associated with microRNA [6]. However, either their prediction performance or the range of their application is still not satisfactory. In addition, some problems in this domain haven't been extensively studied, such as predicting drug resistance induced by microRNAs [7]. The main challenges to consider and integrate multiple relevant node information data for predicting the links in different types of microRNA-mediated biomolecular networks are manifold, including the limited number of training samples, the high dimension and incompleteness of node attributes, and etc [8].

There are problems with the combination of known network structural data and multiple node information data for link prediction. In this thesis, three challenging problems for the task of link prediction are concerned, which are related to data incompleteness of graph-based node information, high dimension of node attributes and data integration of node similarity, respectively [9]. We analyze and address them with three application related to microRNA-mediated biomolecular networks. Multiple biological data are leveraged to solve three relation learning problems regarding the microRNA's regulation mechanism, pharmacotherapy affect, and induced disease. Specifically, the link prediction problems in three different microRNA-mediated biomolecular networks,

i.e., lncRNA-microRNA interaction, microRNA-drug resistance association, microRNA-disease association, are concerned. Hence, we propose different computational solutions to different issues posed by the challenges in these problems. Generally, we consider these three networks as *Attributed Graphs*, which contains vertices, undirected and unlabeled edges, and node attribute. The data type of node attribute in graphs for different prediction problems would be different. For example, it can be numerical vectors, binary features, and edges of subgraph [10]. Their type of links can also be different. For example, the edges for lncRNA-microRNA interaction are weighted while those for microRNA-disease association are unweighted.

The rest of this section is organized as the following. In Section 1.1, the background information about current state of bioinformatic research of microRNA-mediated biomolecular networks, which also motivates us to correspondingly propose effective computational models, is given. In Section 1.2, the challenges posed by the link prediction tasks, are illustrated. Then, we correspondingly explain how to formulate a learning problem to address the challenge. In Section 1.3, the algorithms that may address the challenges are introduced. In Section 1.4, we give the organization of the thesis.

1.1. Motivation

With the advent of high through-put techniques, availability of large-scale experimental data for cell biology is enabling computational methods to systematically model the behavior of biomolecular networks that microRNAs are involves [11]. As a

complement to biological tests and clinical trials, which are expensive and time-consuming, computational approaches have been proved to be able to effectively predict new biomolecular relations or the functions and functional clusters of microRNAs that are testable [5, 12]. Their prediction performance could be very promising especially if the topological relations in networks and the information of multiple molecular attributes success to be considered collectively in prediction. Multiple types of microRNA data have been being accumulated from large-scale experimental technologies, which reflect their physicochemical property (e.g. sequencing information) [13], biological characteristics (e.g. expression profiles) [14] and the interplay between them and other biomolecules (e.g. microRNA-mRNA interactions) [15]. In their most basic abstraction level, microRNA-mediated biomolecular networks can be represented as mathematical graphs, using nodes to represent biological entities (e.g. microRNA, drug and disease), edges to represent their various types of interactions/association, and node attributes to represent different side information (e.g. microRNA functional annotation and disease semantical descriptors). This representation of biomolecular networks as graphs makes it possible to systematically investigate the network topology and functions of biomolecules using conventional graph-theoretical concepts and methods. Several relational learning problems in this domain have been extensively studied. For example, there have been more than 10 kinds of computational approaches proposed for predicting the interactions between microRNA and mRNA [4]. Meanwhile, however, some new research topics that have been emerging in this domain, e.g., the interaction of

microRNA to other kinds of noncoding RNA, remains to be investigated in a bioinformatic way [16].

In this thesis, the first issue that we are interested in the link prediction task is about the data incompleteness in multiple graphs. Its corresponding work that we concerned is to predict large-scale microRNA-disease associations considering useful side information [17]. To date, even though an increasing number of microRNA-disease association have been identified by clinical research, its number is still far from enough to describe the landscape of the effect of microRNA on disease development [18]. There are a number of public databases offer the ground true data of microRNA-disease association that are experiment confirmed, which, for example, include MiRCancer [19], MiR2Disease [20], HMDD [21], DbDEMC [22], OncomiRDB [23] and OncomiRdbB [24]. This offers the information resource for training computational models for prediction, as many researchers aim to do. There have been up to 20 kinds of computational methods proposed to predict microRNA-disease association [25]. Although the algorithms they use are different, their basic assumption is similar, which is that similar microRNA tend to be involved in similar disease, such that how to measure the similarities of microRNAs and diseases is important for these algorithms [6]. Thus, they generally adopt similarity-based algorithms to infer new microRNA-disease association. There are also several public databases collecting comprehensive intrinsic information of microRNA, such as miRbase [26], miRGator [27], miRGen [28] and IntmiR [29], and some ones collecting microRNA-related biological interactions, such as miReg [30], miRTarBase [5], miRecords [31], miRWalk [32] and TarBase [3]. However, in most

published works in this domain, there are only two kinds of data previously used as side information for mearing the functional similarity of microRNAs. One is that calculated by Wang et al. in 2010 and the other is the sequential similarity [18]. To improve the prediction performance, it is needed to consider another type of side information and use a more effective training model to learn the complex nonlinear relation of microRNA features.

The second research problem in this thesis is about how to the high dimensional node attributes. We face this problem when predicting the association between microRNA and drug resistance. Accumulated evidence shows that, in general, the failure of drug treatment is closely associated with the expression of microRNAs [7, 8]. As the gene expression of microRNA varies from person to person, the treatment for the patients having the same disease could be different [33]. Therefore, it is important for personalized treatment to figure out which microRNA can negatively influent the drug effect for a specific type of drug [34]. In addition, it can also offer great insights into the pharmacology of old drugs, which is important for guiding drug reposition [35]. Despite its importance, no computational approach has been proposed to predict drug resistance induced by microRNA on a large scale. Recently, the ncDR database has been released offering 5864 relationships between drug compounds and ncRNAs [7]. This offers an opportunity for researchers to use computational methods based on supervised learning to predict unexplored associations between microRNA and drug resistance. In addition, the property information of drugs is well documented now in public database such like DrugBank [36], which makes it possible to further improve

the prediction performance if the side information of microRNA and drug could be effectively considered.

The third concerned problem is learning integrated node feature from multisource similarity, which pose a major challenge for predicting the co-regulation relationship between microRNA and lncRNA, both of which are major research objectives in the field of noncoding RNA research [12, 37, 38]. Many works have been done to annotate the biological functions of microRNA and lncRNA, which are, however, isolated from each other [39-41]. It is now widely accepted that, as two key biomolecules in ceRNA regulation network, lncRNA and microRNA are co-regulated, commonly forming a complex mechanism for gene regulation. lncRNA regulates microRNA function by acting as endogenous sponges to regulate gene expression meanwhile microRNA binds and regulates lncRNA stability [42]. LncRNA-microRNA interaction is gaining research importance especially when investigating the regulation mechanisms in various types of diseases. The number of known lncRNA-microRNA interactions is still limited. Some existing microRNA-target prediction algorithms, e.g., TargetScan [43], that are based on sequential matching can make prediction of any kind of RNA target for microRNA, including lncRNA. However, they were initially proposed for predicting the microRNA-mRNA interaction and the rules they adopted are not applicable for lncRNA-microRNA interaction. Therefore, they are not suitable for link prediction of lncRNA-microRNA interaction. It is an urgent need to propose more effective compactional methods for a more reliable prediction of lncRNA-microRNA interaction.

1.2. Problem statement for link prediction

Considering that the challenges of the three concerned prediction tasks are different, we correspondingly formulate them as three different link prediction problems. Basically, they all are to predict new links in undirected bipartite attributed graphs. However, the data types of their node attributes are different.

To solve the first concerned prediction task (i.e. microRNA-disease association prediction), exiting computational approaches do not consider the incompleteness of side information [25]. In addition, the co-regulation between lncRNA and microRNA, although important, hasn't been considered in the previous studies [6]. In this thesis, we use the lncRNA-microRNA interaction network as side information to consider the interrelationship of microRNA in the mechanisms of different diseases. Therefore, the main challenge of this task is to deal with the incompleteness of the side information graph (i.e. lncRNA-microRNA interaction network) when training a prediction model. To address this, we formulate this prediction problem as a multi-task one, in which link predictions on two biomolecular networks can be simultaneously implemented.

The main difficulty that we face in the link prediction task for the second concerned biomolecular network (i.e. microRNA-drug resistance association) is associated the high dimension of drug structural information data. Using traditional statistical methods (e.g. Jaccard index) to deal with the high-dimension drug structural data for drug similarity would be ineffective, since too much information would be lost in such a pre-processing process. To address this issue, we formulate this problem as an end-to-end

learning problem, in which the high-dimension side information data can be directly used as inputs of the prediction model.

The third link prediction problem is about lncRNA-microRNA interaction. Different from the conventional sequential matching approaches, we consider the interrelationship of different types of lncRNA/microRNA. The main challenge of this task is to obtain the effective similarity of lncRNA/microRNA regarding to their co-regulation pattern. We construct several types of lncRNA/microRNA similarity and found out that the best one is closely related to the expression profiles. We therefore solve this task problem by formulating it as a link prediction on a bipartite graph where the similarity for two types of nodes is given. To improve the prediction performance, we also investigate how different kinds of similarity can be effectively integrated into one in another work.

1.3. An overview of solutions

Given the challenges and the problem formulation mentioned, we correspondingly propose different algorithms to perform the tasks of link prediction associated with microRNA-mediated biomolecular networks.

First, we propose MVMTMDA, an algorithm which can learn feature embeddings of nodes across different graphs. It was applied to learn a novel representation for microRNA function and to predict new associations between microRNA and complex human diseases. We collected microRNA-lncRNA interaction that are experimentally

confirmed to predict new microRNA-disease associations. Both of these two kinds of data can be represented as two graphs which microRNA are jointly involved. Though they are closely related regarding to their biological meanings, they both are incomplete and thus it is unsuitable to directly use one to predict the links of the another. To solve this problem, MVMTMDA adopts multi-task learning to calculate the possibility of association for each node pair in two graphs. Specifically, it learns feature embeddings for microRNA nodes from two graphs using multi-view learning and performs prediction via multi-task learning. In addition, its deep-learning model structure enable our algorithm applicable to large-scale input graph data with an end-to-end prediction.

Second, we propose GCMDR[44], which is an algorithm for predicting new kinds of drug resistance in which a specific microRNA is involved. Considering the dimension of side information of drug/microRNA, e.g. drug second structure fingerprint, could be as high as thousands, this algorithm adopts graph convolution operator such that can perform prediction without any data pre-processing. Through using an auto-encoder deep learning model structure, GCMDR learns embedding features for microRNA and drug-disease pairs, which leads to a significant improvement on prediction performance.

Third, we propose two algorithms for predicting new interaction between lncRNA and microRNA, which are EPLMI [12] and LMNLMI [45]. In the work of EPLMI, we investigate three types of side information for lncRNA and microRNA in order to depict their comprehensive property. We statically analyzed the latent pattern of known lncRNA-microRNA interaction network and found that expression profile data is the

most effective information for describing microRNA and lncRNA regarding to their interaction. EPLMI is based on a two-way diffusion method propagating labels of node through the networks. LMNLMI is another kind of algorithm, which improve the performance of EPLMI. Different from EPLMI which use just single type of side information for prediction, LMNLMI proposes a similarity fusion method to learn a integrated similarities for microRNA and lncRNA using all kinds of information.

The proposed algorithms have been used in different applications associated with microRNA-mediated biomolecular networks. According to the different challenges posed by different tasks, we concern about how to reformulate the problems and correspondingly develop algorithms to address them. All experiments we did are conducted based on real data sets collected from public databases. The experimental results prove the superior performance of these proposed algorithms than state-of-art approaches.

1.4. Thesis organization

To illustrate the proposed prediction models and present our works, the rest of this thesis is organized as the followings.

In section 2, we review the previous works that are related to the issues in link prediction problems we concern, which include conventional link prediction algorithms considering graph topology information, feature embedding techniques and popular link prediction methods used in bioinformatic. We categorize them into different classes.

In section 3, we give the biological background about the association between disease and lncRNA-microRNA co-regulation and also the research status in this field. We present the main challenges in solving this problem and drawbacks of the previous works, which motivate us to propose a new kind of prediction model. Then, we illustrate how the MVMTMDA model the problem and how it is designed. Finally, we present the experiments we conducted with the results, which can test the efficiency of the proposed model.

In section 4, we first explain the biological background about the association of microRNA's expression and the drug resistance in disease treatment. Then, we discuss about the obstacles in predicting such associations and then show how the proposed GCMDR model formulate it. Comparison experiments were conducted to illustrate the effectiveness of GCMDR model. We present the experimental setting as well as the results at the end.

In section 5, we introduce the biological background under which we try to propose a new method for predicting lncRNA-microRNA interactions. Then, we present the research status in this field and illustrate the challenges faced by the previous works. We explain why we formulate the prediction task for lncRNA-microRNA as a link prediction one in a bipartite attributed graph. We describe and analyze the side information that we collected for depicting the biological property of microRNA and lncRNA. The computational pipelines for two proposed algorithms, i.e. EPLMI and LMNLMI are introduced in detail. To evaluate the prediction performance of the

proposed methods, we show the experimental setting and results, which are conducted on the same data set.

At last, in section 6, we summarize the contributions of the thesis and propose future works.

2. OVERVIEW OF THE RELATED WORK

The proliferation of data that can be represented as graphs have created new opportunities for data analytics in various domains including computational biology, social network analytics, knowledge graph etc. Unsurprisingly, there have been a considerable number of algorithms proposed for link prediction. Though their aims are the same, which is to measure pairs of unconnected nodes in graphs with scores proportional to the likelihood of the existence of in-between links, the methodologies they use could be quite different according to the assumption, input data type considered, problem formulation and particular application field. In this section, state-of-the-art methods for link prediction are introduced categorically.

2.1. Graph Topology-based algorithm for link prediction

As part of the recent surge of research on data mining and machine learning, a considerable amount of attention has been devoted to the computational analysis of graph data. To predict links that are not yet observed but most likely to exist in graphs, there have been an array of methods proposed for link prediction. In general, link

prediction algorithms assign a connection weight sore to each pair of nodes, considering the information data of input graph and node attributes. By ranking the list of weight source in decreasing order, the most potential real links can be predicted among the unobserved node pairs. In this section, we categorize and survey state-of-the-art link prediction techniques previously proposed.

2.1.1. Methods based on node neighborhoods

As a natural intuition regarding to real application, nodes sharing more neighbors tend to construct links. For examples, in social networks, authors who have many colleagues in common are more likely to contact themselves. Therefore, some link prediction algorithms consider the local topological information of nodes in graphs. For a node x, we denote the set of its neighbors N(x). Given two nodes x and y in a graph, there are a number of approaches proposed to measure their similarity based on their neighbors, i.e. N(x) and N(y). For example, Newman [46]verified the correlation between x and y by considering their Common Neighbors, $|N(x) \cap N(y)|$. Josipa [47] proposed another similarity metric of this kind, Jaccard's coefficient, which can be formulated as $|N(x) \cap N(y)|/|N(x) \cup N(y)|$. Some works [48] assume the probability that a new link containing node x is proportional to $|N(x)|$. Some works also extended this assumption on the basis of empirical evidence, considering this probability is associated with preferential attachment of nodes $|N(x)| \cdot |N(y)|$.

2.1.2. Methods based on the Paths

Different from the methods considering the local topological information of node

neighborhoods, a number of methods calculate the measurement of paths to consider the more general information of graph topology. The most basic method of this category is to calculate the length of shortest path between nodes x and y as their correlation score [49]. However, it can cause extremely computational cost when being applied on huge graphs. Instead of computing the shortest-path distance, a number of methods implicitly consider the ensemble of all paths between a pair of nodes. For example, Katz propose a measure which sums over the collection of paths, exponentially damped by length to count short paths heavily [50]. The Katz measure can be formulated as $Katz(x,y) := \sum_{l=1}^{\infty} \beta^l \cdot \left|paths_{x,y}^{\langle l \rangle}\right|$, where $paths_{x,y}^{\langle l \rangle}$ denote the set of all length-l paths from x to y and β<0 is a parameter. One solution for its score matrix is given by $(I-\beta A)^{-1}-I$, where A is the adjacency matrix of the graph. FriendLink [51] metric is another path-based method to calculate the similarity between two nodes, x and y. Its definition is $FL(x,y) := \sum_{i=1}^{l} \frac{1}{i-1} \cdot \frac{|paths_{x,y}^{i}|}{\prod_{j=1}^{i}(n-j)}$, assuming that nodes in a graph can use all the paths between them to form connection. Other variant algorithms include Local Path [52], Relation Strength Similarity [53], Vertex Collocation Profile [54], etc.

There are also some methods based on the technique of random walk [55]. Two typical examples of this type are Hitting Time [56] and PageRank [1]. Random walk is an operator that walks on the input graph starting at a node x and iteratively moves to a neighbor node at random. The measure of Hitting Time for nodes x and y is the expected number of iterations required by a random walk to move from x to y. However, Hitting Time is sensitive dependence to the parts of graph far away from x and y, even if x and y are connected with short paths. One way to solve this problem is to let random walk

restart periodically from x, such that parts of distant paths in graph are not likely to be explored by the random walk. On this basis, PageRank algorithm was proposed for Web pages. The algorithm of SimRank [57]considers that two nodes are similar to the extent that they are connected with similar neighbors. By specifically defining the similarity(x, x) =1, the definition of SimRank measure is recursive and can be formulated as $\text{similarity}(x, y) := \gamma \sum_{a \in N(x)} \sum_{b \in N(y)} similarity(a, b) / |N(x)| \cdot |N(y)|$.

2.2. Graph embedding for link prediction

The data amount today is becoming extremely huge in field of data analytics, though effective regarding prediction accuracy, the above strategies which are proposed to directly perform prediction suffer the high computation and space cost. An emerging technique called Graph Embedding provides an efficient solution to solve the graph analytics problem [58]. Specifically, it converts a graph into a low dimensional space where the information of graph topology and node attribute can be maximumly preserved. For example, given two nodes connected in a same graph with their attribute, graph embedding algorithms can learn their features that close to each other in a low dimensional space. The learned embedding features could be a representation for nodes, vectors or a whole graph and thus conducive to downstream prediction problem such as graph classification, link prediction, node prediction, graph clustering and etc. Graph embedding can extract features for each data entity to represent their roles in the networks by comprehensively considering the network topology and the node attributes. In this section, we categorize and introduce existing algorithms based on embedding

techniques for link prediction.

2.2.1. Graph embedding based on matrix factorization

By representing a graph in a form of matrix (e.g. adjacent matrix), matrix factorization-based graph embedding normally factorize this matrix along with other graph property information (e.g. node attributes) to obtain embedding features for nodes or vectors. In most cases, the inputs of such kind of algorithm are matrixes representing the graph topology and high dimensional node attributes and the outputs are the feature matrixes for nodes. Therefore, different kinds of matrix factorization techniques have been proposed to tackle this problem which can be treated as a structure-preserving dimensionality reduction problem. In general, they can be classified into two types. One is based on factorization on graph Laplacian eigenmaps and the other is on node proximity matrix. The former kind assumes that nodes close in a graph should have similar embedding features and thus impose larger penalty if those similar nodes are embedded far apart. For example, [59] optimize node embedding y using the objective function with Laplacian eigenmaps: $y^* = \arg\min \sum_{i \neq j} (y_i - y_j{}^2 W_{ij}) = \arg\min y^T L y$, where W_{ij} is defined as the similarity between node i and j in a graph, and L is the graph Laplacian of W. To remove an arbitrary scaling factor in L, [60] further constrains $y^T D y$ to be 1, where D is the diagonal matrix ($D_{ii} = \sum_{j \neq i} W_{ij}$). The corresponding objective function is thus reduced to be $y^* = \arg\min y^T W y / y^T D y$. To solve the cold start problem, [61] use a linear function $y = X^T a$ for embedding so that the algorithm can be applied to those new nodes with feature X. It turns to solve the

objective function $y^* = \arg\min a^T XWX^T a / a^T XDX^T a$. The differences of the methods of this kind mainly lie in the way to calculate the node similarity matrix *W*. The choices include Euclidean distance between node features [62], k-Nearest Neighbor (KNN) [63], anchor graph technique [64], local regression model [65], local spline regression [66], principal component analysis (PCA) [67] and semidefinite programming (SDP) [68]. In addition to solving the Laplacian eigenmaps, another solution is to factorize node proximity matrix. Methods of the second kind assume that it is feasible to approximate node proximity in low-dimensional space using matrix factorization. Generally, given the matrix of node proximity *W*, their objective functions are $\min\|W - YY^{cT}\|$, where *Y* is node embedding and Y^c is the one for context nodes [69]. Solutions to find rank-d approximation of *W* can be based on singular value decomposition (SVD) [70, 71], regularized Gaussian matrix factorization [72], low-rank matrix factorization [73] etc.

2.2.2. Graph embedding based on deep learning

With the advent of deep learning techniques, a considerable number of prediction models for graph analytics have been designed with deep neural network structures. Most of them are based on graph embedding, which can be categorized into three types according to the techniques they used: random walk, Autoencoder and Convolutional Neural Network.

In the first category, the input is paths sampled from a graph, based on which the deep learning methods are then applied for embedding. For example, DeepWalk [74] first

samples a set of paths from the input graph using truncated random walk [55] and then applies a neural language model (SkipGram [75]) on the path set to maximize the probability of observing nodes' neighborhood close in the embedding space. Along this direction, different kinds of deep learning techniques have been proposed for graph embedding following random walk, which include long-short term memory (LSTM) [76], GRU [77], DCNN [78] and etc.

The second category adopts the autoencoder model aiming to reconstructing the input graph by minimizing the reconstruction error of the input and output [79]. In encoder and decoder component, the structure contains neural network layers with multiple nonlinear functions. The encoder maps the input graph into embedding space in which the decoder then maps the embedding features to reconstruction space. In general, the embedding space has low dimension such that compacted representation can be obtained. This idea is similar to the solutions that factorize the node proximity matrix. Examples include SDNE [80], DNGR[81], SAE [82].

Model structures of the third category are based on convolutional neural network (CNN) and its variants. CNN model was originally signed for figure data which is in Euclidean domain. Some attempts have been made to reformat the graph input to fit the input of original CNN model. For example, [83] adopts graph labelling to learn neighborhood representation of nodes as inputs of CNN. On the other hand, some works recently attempt to redefine the graph convolution-like operation using spectral graph theory. Representative algorithms of this kind include [9, 84-86]. Spectral-based methods have

a solid mathematical foundation in graph signal processing. They assume graphs to be undirected. The graph convolution operation is proposed based on the normalized graph Laplacian matrix, which is a mathematical representation of an undirected graph. On the other hand, analogous to the convolutional operation of a conventional CNN on an image, spatial-based methods define graph convolutions based on a node's spatial relations. The main differences of these two graph convolutions lies in whether the operation is conducted on spatial domain or spectral domain.

2.3. Link prediction in bioinformatics

Graph data naturally exist in a wide diversity of biological study, e.g., protein-protein interaction, gene-disease association, drug-protein interaction etc. Analyzing these graphs can provide insight into making good use of the information pattern hidden the graph, and thus attract increasing attention in the field of bioinformatics. Different kinds of prediction models have been proposed for graph analytics adjusting to the different kinds of biological data, mostly heterogeneous biological networks. In this section, we categorize the most popular kinds of prediction model in bioinformatics. In general, as the first step in the computational pipeline of these methods, the similarity of the research entity is needed to be computed. For example, [87] uses the graphs of MeSH to calculate the disease semantic similarity. [88] uses Gaussian interaction profile kernel to construct microRNA similarity networks. After obtaining the similarity matrix for nodes in graph, as the second part in the computational pipeline, different techniques have been proposed to perform prediction, which can be classified into two categories

according to whether local or global graph information is considered. We here take the existing studies of microRNA-disease association prediction as examples. The methods considering local information generally adopts techniques that enjoy a low computation cost, such like cumulative hypergeometric distribution [89], label propagation [90], KNN [91] and regularized least squares [92]. On the other hand, a considerable number of methods have been used to consider the general information, including random walk [93], various types of path-based inference methods [94, 95], restricted Bolzmann machine [87], support vector machine [96, 97], matrix completion[98] etc.

3. MVMTMDA - A MULTI-VIEW MULTI-TASK LEARNING ALGORITHM FOR PREDICTING MICRORNA-DISEASE ASSOCIATIONS

3.1. Background

MicroRNAs and lncRNAs have been found involved in transcriptional and post-transcriptional processes, forming a gene expression program that all eukaryotic cells rely on [16]. MicroRNAs are ~22nt ncRNAs and they generally bind to the 3'UTR of the mRNA imperfectly. In most cases, this can lead to translational inhibition or degradation of its target mRNA. Although much effort has focused on the functions and biogenies of microRNAs, lncRNAs are gaining prominence as they take up the largest portion of mammalian non-coding transcriptome. It has recently been found to serve the role of critical epigenetic regulators of gene expression [99]. Most diseases are

frequently associated with alteration of the transcriptome, and such an altered transcription pattern has recently been found to not just be restricted to the protein-coding RNAs aberrantly expressed but also to dysregulation of the expression of microRNAs and lncRNAs. As a result, much effort is currently being made to characterize those lncRNAs and microRNAs that interfere with gene expression and signaling pathways at various stages of disease development.

Recently, there has been an increasing body of experimental evidence that shows that, through a sophisticated and multi-layered mode of regulation noncoding RNA, including lncRNA and microRNA, can influence every aspect of normal tissue physiology [16]. Recently, the competitive endogenous RNA (ceRNA) hypothesis [100] has gained substantial attention as it unifies all the hypotheses about the general mechanism of the intricate interplay among diverse RNA species. Specifically, it proposes that lncRNAs that share specific MREs (microRNA response elements) communicate with and co-regulate each other by competing for binding to the shared microRNAs. Considering that both lncRNA and microRNA are key regulators that control cellular processes, and that they interact with each other to fine-tune gene expression, knowledge of the mechanisms by which they cooperate is the first step towards understanding the functions that they exert in disease processes. Unfortunately, in spite of its importance, little is known about the co-regulation between lncRNA and microRNA in disease processes.

However, with the advent of high-throughput sequencing techniques, more and more lncRNAs and microRNAs have been identified to be involved in the development of diverse diseases, which include cancers, acting as oncogenes or as tumor suppressors [17]. Both lncRNA and microRNAs are now routinely used as biomarkers in disease diagnosis and treatment. Much progress has also been made towards their use as molecular targets for new drugs. This promising trend depends largely on our understanding of the associations between lncRNA or microRNA and a diversity of different diseases.

Recently, with the advances in analytical methods including circulation, genetic, epigenetic, microRNA-target and tissue-expression assays, several databases, such as HMDD [17] and miR2Disease [20], have been established to allow data related to the relationship between lncRNA/microRNA and different diseases to be publicly accessible. Unfortunately, as the assays are time consuming and tedious, the data that have been collected so far are still relatively, focusing only on a few key noncoding RNAs rather than their contextual regulation network. In addition, it can be difficult to integrate the data in the databases together to form a complete regulation network due to their sparsity in number and their being from different bioassay-based research.

In recent years, there has been increasing research interest to exploit LMI (lncRNA-microRNA interaction) in studies related to various complex human diseases [101] such as colorectal cancer [38, 102], cervical squamous cell carcinoma (CESC) [37] and heart failure , etc. Rather than investigating into the signaling pathways of a few types of

noncoding RNAs, these studies consider transcriptome-wide regulation that involves both microRNAs and lncRNAs cooperating together. However, it should be noted that, as information about lncRNA-microRNA regulation network is not available from existing databases, current research in this area is mainly based on sequence-based microRNA target-prediction algorithms, such as miRWalk [32], Cytoscape [103] and TAM [104]. These algorithms are used to construct a predicted LMI network so that they can be used to predict pathogenic lncRNA-microRNA co-regulations. However, as pointed out by some studies, most existing microRNA target-prediction algorithms predict too many false positives and the LMI networks constructed based on the results that these algorithms predict would therefore be unreliable [105]. Although ground-truth data of LMI may help us understand the important regulation functions of non-coding RNA (ncRNA) thereby deciphering the complex ncRNA regulation network in the pathology of diseases, finding out the relationship between LMIs and the diseases that they are associated with is difficult.

As it is slow and tedious to perform laboratory experiments, relying on computational approaches can allow potential candidates for experimental confirmation to be quickly identified by better integrating prior information from different relevant studies much faster and with much lower costs. Towards this goal, a number of computational tools have been developed for computer-aided ncRNA biomarker discovery. As reviewed in [6, 106], most existing methods in this domain rely on the basic assumption that microRNAs that are similar tend to be involved in diseases that have similar pathological characteristics. While this assumption seems to be very reasonable, it

should be noted that how microRNA similarity should be defined is a complex and open problem. Different metrices for microRNA similarity have been proposed using different side information and statistical metrices such as Pearson correlation coefficients, cosine similarity and Euclidean distance [6]. However, as the features in the feature vectors of the side information may not be linearly dependent, these metrices may not be able to capture the complex relationship between two lncRNA/microRNA. In addition to this problem, the data of side information, such as the LMIs, are quite limited in amount and are incomplete. Due to the missing data, microRNA similarity cannot be determined accurately and for this reason, MDAs cannot be predicted accurately. Hence, in order to improve prediction accuracy, there is a need to learn an effective feature representation for microRNA and lncRNA.

There has recently been an increasing number of computational tools proposed to predict MDA. Many of these tools do not take into consideration the incompleteness of information about what raw features of microRNAs can best be used for prediction. Also, these tools determine functional similarity for microRNAs based on data sources that are not reliable [6, 106]. For example, the functional similarity score matrix (http://www.cuilab.cn/files/images/cuilab/misim.zip) released by Wang et al. [18] was obtained using a computational model developed with a data set which has not been continually updated, and as such, prediction of the association relationships are not reliable. One other limitation with existing tools is related to the statistical methods that they use to compute the microRNA similarity scores. As explained above, they are too simple to capture the complex correlation relationship among microRNAs. For example,

the use of Gaussian kernel measure or linear Euclidean distance, which are widely used, do not capture dependence of features in microRNA feature vectors [106]. In summary, the choice of data sources and the way side information is integrated do not provide current computation methods the best tools to predicting MDAs most accurately.

3.2. MVMTMDA in details

To develop better methods for this purpose, we take the co-regulation between lncRNA and microRNA into account when predicting new microRNA biomarkers. Based on the assumption that the patterns of lncRNA-microRNA co-regulation can be implied from the network of LMI identified by large-scale CLIP-seq experiments, we developed a computational model to predict MDAs on a transcriptome-wide scale by introducing known LMIs.

3.2.1. Data collection

The data we used in this work include experimentally-validated lncRNA-microRNA interactions and MDAs. There are several public databases providing such two types of data resource. In order to obtain the up-to-date data resource, we collect the datasets from lncRNASNP v2.0 [107] and HMDD v3.0 [17], which of both have been recently updated within a year.

lncRNASNPv2.0 database (http://bioinfo.life.hust.edu.cn/lncRNASNP) integrates the data from starBase v3.0 [3] database (http://starbase.sysu.edu.cn/) providing comprehensive knowledge on lncRNAs. It records 45329 LMIs between 3521 types of

lncRNA and 276 types of microRNAs. HMDD v3.0 database (http://www.cuilab.cn/hmdd) provides 18732 MDAs between 874 types of diseases and 1207 types of microRNAs.

We discard the redundant data and manually match the ids of microRNA in lncRNASNP v2.0 datbase and those in HMDD v3.0 database. As a result, the LMI dataset we collected has 10465 LMIs between 541 lncRNAs and 268 microRNAs. Based on the 268 types of microRNAs whose ids are successfully matched to HMDD database, we collect 11253 MDAs covering 799 types of diseases.

3.2.2. Problem statement

In this work, we propose MVMTMDA to predict MDAs considering the co-regulation of lncRNA and microRNA. As mentioned in the Introduction section, one challenge of our work is to solve the problem of incompleteness and sparsity of LMI networks. To this end, we introduce what we call *multi-task learning* when we design our model. Based on multi-task learning, LMIs and MDA are simultaneously predicted. Considering that both known networks of LMI and MDA are far less than complete and that the information contained in these networks are complementary to each other, we believe, therefore, that prediction of new links in one network can be made based on the other ones. These predictions are also mutually beneficially. An accurate link prediction in the LMI network, for example, can provide useful information for MDA predictions to be made more accurately and vice versa.

Another challenge of our prediction task lies in the development of a similarity measure between lncRNA/microRNA in a lncRNA-microRNA-disease network. Due to complexity of their synergistic effects, such a measure would be highly complex as well. To tackle the problem, we propose to learn embedding features for lncRNA and microRNA from the LMI and MDA networks and this can be defined as a multi-view learning problem. We consider the functional roles of a given microRNA to have two heterogenous representations on LMI and MDA networks, respectively, with each network having a different view. The key to tackling the multi-view learning problem is to effectively exploit the diversity and consistency of multi-view data of the networks of LMI and MDA, which consequently identifies the feature dimensions in which the characteristics of the original data could be retained.

Suppose there are N_m types of microRNAs $\mathcal{M} = \{m_1, \dots, m_{N_m}\}$, N_d types of diseases $\mathcal{D} = \{d_1, \dots, d_{N_d}\}$ and N_l types of lncRNAs $\mathcal{L} = \{\ell_1, \dots, \ell_{N_l}\}$. Let $\mathcal{X} \in \mathbb{R}^{N_d \times N_m}$ and $\mathcal{S} \in \mathbb{R}^{N_l \times N_m}$ denote the adjacent matrixes of known MDA and LMI network, respectively. Based on the datasets we collected, $\mathcal{X}$ and $\mathcal{S}$ are constructed as,

$$\mathcal{X}_{ij} = \begin{cases} 1, & if\ d_i\ and\ disease\ m_j\ is\ known\ to\ be\ connected \\ 0, & otherwise \end{cases} \quad (1)$$

$$\mathcal{S}_{ij} = \begin{cases} 1, & if\ \ell_i\ and\ disease\ m_j\ is\ known\ to\ be\ connected \\ 0, & otherwise \end{cases} \quad (2)$$

We formulate prediction task for MDAs as the problem of simultaneously estimating the value of each unobserved entry in $\mathcal{X}$ and $\mathcal{S}$. It is assumed that there is an

underlying model which can be constructed to generate all interaction possibility for each pair of MDA/LMI as follows.

$$\widehat{\mathcal{X}}_{ij} = F_x(d_i, m_j, \mathcal{S}|\Theta_{\mathcal{X}}) \tag{3}$$

$$\hat{\mathcal{S}}_{ij} = F_{\mathcal{S}}(\ell_i, m_j, \mathcal{X}|\Theta_{\mathcal{S}}) \tag{4}$$

where $\widehat{\mathcal{X}}_{ij}$ and $\hat{\mathcal{S}}_{ij}$ denote the predicted score of association $\mathcal{X}_{ij}$ between disease d_i and microRNA m_j and interaction $\mathcal{S}_{ij}$ between lncRNA ℓ_i and microRNA m_j, respectively; $\Theta_{\mathcal{X}}$ and $\Theta_{\mathcal{S}}$ denote the model parameters; F_x and $F_{\mathcal{S}}$ denote the functions that map the model parameters to predicted scores. As the outputs of F_x and $F_{\mathcal{S}}$ are also the inputs for each other, we adopt a co-training optimization method to train the models. We introduce latent factor model (LFM) to build functions F_x and $F_{\mathcal{S}}$ applying dot product as,

$$\widehat{\mathcal{X}}_{ij} = F_x^{LFM}\left(d_i, m_j, \mathcal{S}\middle|\Theta_{\mathcal{X}}\right) = p_i^T q_j \tag{5}$$

$$\hat{\mathcal{S}}_{ij} = F_{\mathcal{S}}^{LFM}\left(\ell_i, m_j, \mathcal{S}\middle|\Theta_{\mathcal{S}}\right) = q_i^T r_j \tag{6}$$

where p, q and r denote the latent features for disease, microRNA and lncRNA, respectively. For the purpose of learning the non-linear connection among lncRNA, microRNA and disease, in this work, we propose a method to simultaneously learn the function F_x and $F_{\mathcal{S}}$ with three multi-layer neural networks.

We present a deep neural network to exploit features of lncRNA, microRNA and disease with graph embedding on lncRNA-MDA network. The learned features synthesize the

information in the networks of LMI and MDA and thus is anticipated to comprehensively describe functional role and correlations of lncRNA and microRNAs.

3.2.3. Model structure of MVMTMDA

The proposed model, MVMTLMDA, is designed with a deep structure composed of three neural networks. Different from conventional prediction models for MDA which separate similarity measurement and value prediction, it provides end-to-end solution to handle graph-based raw data to yield the final results without any statistical assumption. Specifically, it learns the hidden features for diseases, microRNAs and lncRNAs via multi-view learning and yields the prediction via multi-task learning (see Figure. 1).

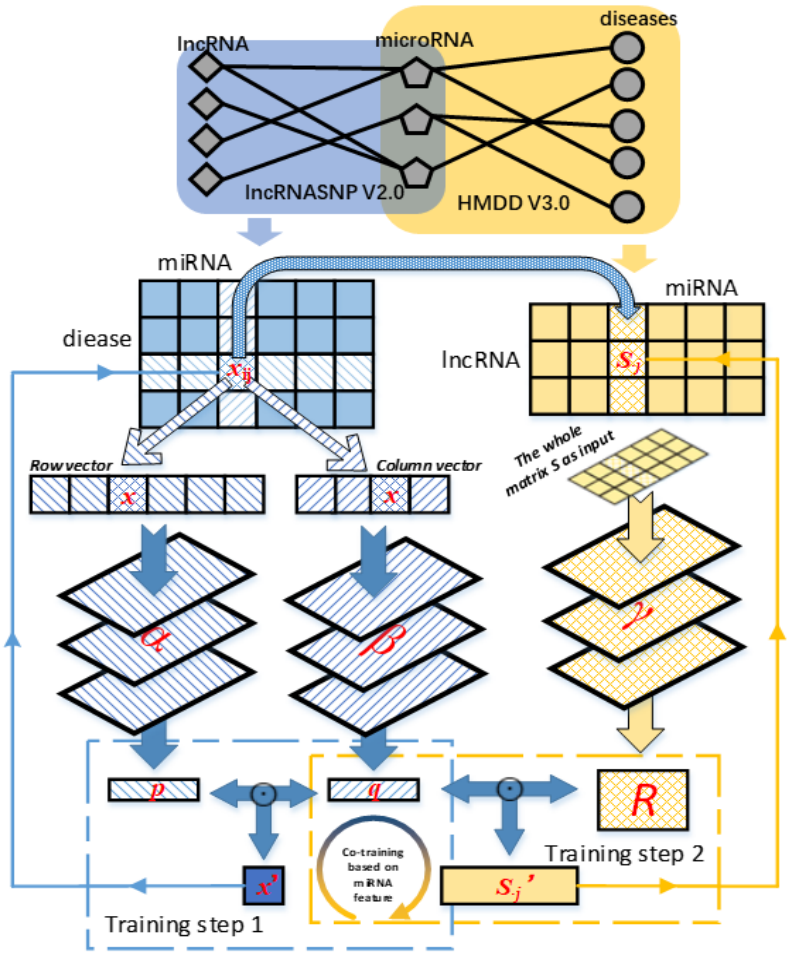

Figure 1 Schematic diagram of multi-view multi-task learning for microRNA-disease association prediction.

To the best of our knowledge, this work is the first attempt to consider the co-regulation between lncRNA and microRNA to predict MDAs. Apart from the prediction improvement from previous models, the contribution of our work lies in some outstanding characters of our method which can be outlined as follows: i) MVMTMDA is able to integrate data from different types of relevant biological network for prediction even if the data is incomplete; ii) it realizes end-to-end training for feature representation from multiple biological networks; iii) it provides a solution to combine the issues of MDA prediction and lncRNA-disease association prediction.

3.2.4. Multiple graph embeddings via multi-view learning

As mentioned in Section 3.2.2, we form two matrix X and S according to Equation 1 and 2. With matrix X and S as inputs, we propose an architecture of three deep neural network to project each types of disease, microRNA and lncRNA into a latent structured space. From the matrix X, each disease d_i is firstly represented as a i-th row vector X_i, which represents the i-th disease's relationship across all microRNAs. Each microRNA m_i is firstly represented as a j-th column vector $X_{\cdot j}$, which represents the j-th microRNA's relationship across all diseases. As shown in Figure 1, the input feature of each type of elements on lncRNA-microRNA-disease network is processed by a single neural network. In each layer of both networks, each input vector is mapped into another vector of a different dimension in a new space. Denote a given input vector of a neural network by x, the output vector by y, the intermediate hidden layers by l_i, i = 1,…,N-1, the weight matrix and bias term of l_i by W_i and b_i. We have

$$\begin{gathered} l_1 = W_1 x \\ l_i = f(W_{i-1} l_{i-1} + b_i), \;\; i = 2, \dots, N-1 \\ y = f(W_N l_{N-1} + b_N) \end{gathered} \tag{7}$$

where the activation function f is here chosen as the ReLU function, $f(x) = \max(0,x)$. In the neural network γ that is built to considering the side information of LMI, the whole adjacent matrix S for LMI network is used as the input. With a given training sample x_{ij}, its aim is to learn the features of all lncRNAs according their known interaction with the j-th microRNA. According to Equation 7, the outputs of neural networks α, β, γ can be respectively formulated as follows:

$$p_i = f_{\theta_N^\alpha}\left(\ldots f_{\theta_3^\alpha}\left(W_{\alpha 2} f_{\theta_2^\alpha}\left(X_{i\cdot}{}^T W_{\alpha 1} + b_{\alpha 1}\right) + b_{\alpha 2}\right) + b_{\alpha 3} \ldots\right) + b_{\alpha N} \quad (8)$$

$$q_j = f_{\theta_N^\beta}\left(\ldots f_{\theta_3^\beta}\left(W_{\beta 2} f_{\theta_2^\beta}\left(X_{\cdot j} W_{\beta 1} + b_{\beta 1}\right) + b_{\beta 2}\right) + b_{\beta 3} \ldots\right) + b_{\beta N} \quad (9)$$

$$D' = f_{\theta_N^\gamma}\left(\ldots f_{\theta_3^\gamma}\left(W_{\gamma 2} f_{\theta_2^\gamma}\left(S W_{\gamma 1} + B_{\gamma 1}\right) + B_{\gamma 2}\right) + B_{\gamma 3} \ldots\right) + B_{\gamma N} \quad (10)$$

Here, $W\alpha_1$, $W_{\beta 1}$ and $W\gamma_1$ are the weight matrix of the first layer in networks α, β, γ, respectively and $b_{\alpha 1}$, $b_{\beta 1}$, $B_{\gamma 1}$ are the corresponding bias terms. $W_{\alpha 2}$, $W_{\beta 2}$,$W_{\gamma 2}$, $b_{\alpha 1}$, $b_{\beta 1}$, and $B_{\gamma 1}$ are for the second layer, and so on. It should be noted that the row dimension D' is N_l, the same as that of input matrix S. $D' = [r_1{}^T, \ldots, r_{N_l}{}^T]^T$ is a matrix stacking the embedding features of all lncRNAs. Based on the embedding features learned from neural networks α, β, γ, we formulate the outputs of our models as follows:

$$\hat{\mathcal{X}}_{ij} = F^{MVMTMDA}\left(X_{i\cdot}, X_{\cdot j} \middle| \Theta_\alpha, \Theta_\beta\right) = cosine\left(p_i, q_j\right) = \frac{p_i{}^T q_j}{\|p_i\| \|q_j\|} \quad (11)$$

$$\hat{\mathcal{S}}_{\cdot j} = [\hat{\mathcal{S}}_{1\mathrm{j}}, \ldots, \hat{\mathcal{S}}_{N_l \mathrm{j}}]^T = F^{MVMTMDA}\left(X_{\cdot j}, S \middle| \Theta_\beta, \Theta_\gamma\right) = \frac{R \cdot q_j{}^T}{\|R\| \|q_j\|} \quad (12)$$

It should be noted that, because of the operation of dot product in Equation 11 and 12, the weight matrixes in the last layers of neural networks should have the same column dimension, assuring that the dimensions of p_i and q_j and the column dimension of R are the same. The embedding feature of microRNA q_j connects the results of neural networks β and γ, and therefore it remains and combines the information of their inputs (i.e. known MDAs and LMIs). As q_j is used to yield the scores for each pair of MDA and LMI, it can effectively represent the biological role of a given microRNA on both networks while training the model by recovering X and S. We consider X and S are

strongly related data that provide two different view for the function of microRNAs and the embedding features yielded from the proposed are basically based on multi-view learning.

3.2.5. Model training via multi-task learning

Based on the outputs yielded by Equation 11 and 12, we define two objective functions for model optimization according to the observed data and unobserved feedback. Each of object functions is corresponding to one prediction task. Considering that the prediction problem is a semi-supervised learning problem with all the training sample that are positive, the objective function is generalized as follow:

$$\mathcal{L} = \sum_{y \in Y^+ \cup Y^-} l(y, \hat{y}) + \lambda \Omega(\Theta) \tag{13}$$

where $l(\cdot)$ denotes a loss function; $\Omega(\Theta)$ is the regularizer for model parameters; Y^+ is the set of positive samples and Y^- is that of negative samples which we adopt negative sampling on the unlabeled microRNA-disease pairs. To train the neural network α and β on the dataset of MDA, the first loss function is and defined with a binary cross-entropy loss as follow.

$$\mathcal{L}_1 = \sum_{(i.j) \in X^+ \cup X^-} \lambda_1 X_{ij} log \hat{X}_{ij} + (1 - \lambda_1)(1 - X_{ij}) \log\big(1 - \hat{X}_{ij}\big) + \sum_{w \in W_\alpha \cup W_\beta} \|w\|_2 + \sum_{b \in b_\alpha \cup b_\beta} \|b\|_2 \tag{14}$$

where w and b denote the parameters in neural network α and β. For training the model on known interactions between lncRNA and microRNA, the loss function for the second step is defined as follow.

$$\mathcal{L}_2 = \sum_{(i.j)\in X^+\cup X^-} MSE\left(S_{.j} - \hat{S}_{.j}\right) + \sum_{w\in W_\beta \cup W_\gamma} \|w\|_2 + \sum_{b\in b_\beta \cup b_\gamma} \|b\|_2 \qquad (15)$$

where $MSE(\cdot)$ denotes the function of mean-square error. The optimization for model training contains two steps based on L_1 and L_2 which are executed alternately. Optimization on function L_1 is basically a point-wise matrix factorization on LMI network while that on function L_2 is a column-wise matrix factorization on MDA network. As the first step of optimization is to predict the scores for the pairs of MDA while the second step is to predict the interaction possibility for lncRNA-microRNA pairs, the proposed model is basically optimized via multi-task learning.

3.2.6. Prediction of lncRNA-disease association with MDA and LMI

Computational tools for predicting disease-associated noncoding RNAs can be mainly categories into two types, lncRNA-disease association prediction and MDA prediction. Despite their close intrinsic relation with respect to the function mechanism of lncRNA and microRNA, little effort has been devoted to combine these two important fields. We here consider the lncRNA-microRNA interaction as a useful bridge to connect these two prediction problems and propose a statistical method to predict lncRNA-disease association based on the results yielded by MVMTMDA. Based on a score matrix of MDA $\hat{X}$predicted by MVMTMDA and the adjacent matrix of LMI, we calculate the p-

value for each pair of lncRNA-disease. Given a lncRNA-disease pair (l_p-d_p), we denote L_m the number of microRNAs associated with l_p in LMI dataset, D_m the number of microRNA associated with d_p in MDA dataset, and M_{ld} the number of microRNAs which simultaneously associated with lncRNA l_p and disease d_p. The p-value for the association between l_p and d_p is defined as follow:

$$p_{l_p \to d_p} = 1 - \sum_{i=0}^{M_{ld}} \frac{\binom{D_m}{i}\binom{N_m - D_m}{L_m - i}}{\binom{N_m}{L_m}} \tag{16}$$

In the datasets that we collected, each type of lncRNA and disease has relation to at least one microRNA, such that the p-value for each lncRNA-disease pair can be calculated using Equation 16. By setting p-value<0.05, we consequently identify 15945 lncRNA-disease associations from totally 432259 lncRNA-disease pairs. To further control the false positive rate of our prediction, we, in addition, conduct false discovery rate (FDR) correction on the computed p-values. The lncRNA-disease pairs with FDR less than 0.05 are considered to have strong positive or negative correlation. As a result, we identify 25076 potential lncRNA-disease association.

3.3. Experiment and analysis

3.3.1. Performance evaluation for MVMTMDA

To evaluate the prediction performance of the proposed model, we used a real dataset involving experimentally-confirmed MDA and LMI and tested accuracy using 2-fold, 5-fold and 10-fold cross validation. Specifically, in *k*-fold (k = 2, 5 and 10) cross

validation, we randomly separate the samples of MDA into *k* roughly equal parts. *k*-1 of them are in turn used as training samples and the rest one is for testing. To quantify the performance in *k*-fold cross validation, we adopt three kinds of criteria, i.e., AUC, HR and NDCG.

In each fold of prediction, we calculate the ranks of testing samples among the unlabeled samples. Those testing samples obtaining a rank higher than the given threshold are considered as positive. Setting different thresholds, we computed the corresponding true positive rates (TPRs, sensitivity) and false positive rates (FPRs, 1-specificity) where sensitivity and specificity are the percentages of testing samples predicted as positive and negative, respectively. Corresponding receiver operating characteristic (ROC) curves are computed by plotting TPR versus FPR and the area under the curves (AUC) is computed. AUC = 0.5 implies a purely random guess and AUC =1 indicates perfect prediction. In addition, we adopt the metrics of HR and NDCG [108]. We used the testing samples and 50 times its number of random unlabeled samples to construct the Ground-truth item set (GT) and truncated the ranked list at 10 for both metrics [108]. As such, the HR intuitively measures the percentage of testing samples in the top-10 list while the NDCG measures the ranking quality which assigns higher scores to hits at top position ranks. For both metrics, larger values indicate better performance.

To avoid any bias caused by the random sample partitioning in cross validation, we repeat the random sampling along with prediction for 20 times. The performance results

of average AUC, best HR and best NDCG yielded by MVMTMDA are listed in Table 1. As larger size of training set would lead to a more accurate prediction, it shows that the prediction accuracy yielded by the proposed model yields increases with the increased number of folds in *k*-fold cross validation. The corresponding ROC curves are shown in Figure 2(a), 2(b) and 2(c) show the HR and NDCG yielded by the proposed model increase rapidly within the first 10 epochs and tend to stabilize after the 20th training epoch.

Table 1 Prediction performance w.r.t. AUC, HG and NDCG using MVMTMDA in k-fold cross validation

CV method	2-fold CV	5-fold CV	10-fold CV
Average AUC	0.8410+/-0.018	0.8512+/-0.012	0.8521+/-0.008
Best HR	0.7196	0.7553	0.7603
Best NDCG	0.4429	0.4895	0.5030

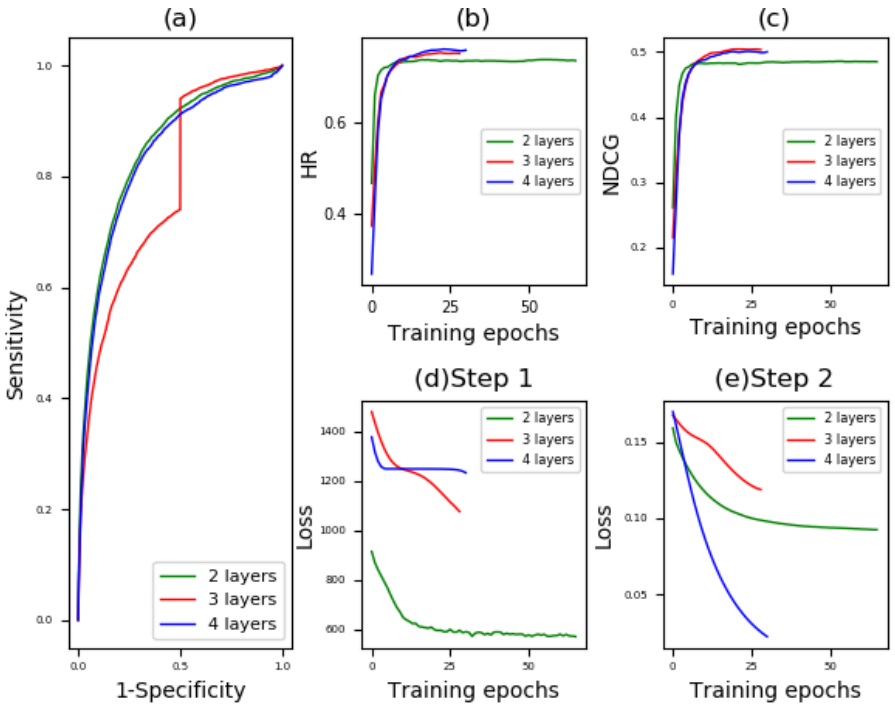

Figure 2 Prediction performance of MVMTMDA: (a)ROC curves yielded by MVMTMDA with 2, 3 and 4 layers; (b) Hit ratio yielded by MVMTMDA with increasing training epochs; (c) NDCG yielded by MVMTMDA with increasing training epochs; (d) the training loss in Equation

3.3.2. Performance evaluation on LMI prediction using MVMTMDA

As mentioned, we consider that the prediction of MDA and that of LMI are mutually beneficial. Given a type of microRNA, its involvement in different diseases offers useful information for predicting its target lncRNA. In this section, we try to use the MDA to predict LMI using MVMTMDA, whose prediction performance is concerned. Specifically, we exchange the matrixes of X and S, with X as the LMI matrix and S as the MDA matrix. We set the model parameters the same as the setting in the above experiment. As a result, predicting LMIs with 2 hidden layers, MVMTMDA yielded average AUC of 0.8747+/-0.018, 0.9014+/-0.012 and 0.9037+/-0.011 in 2-fold, 5-fold and 10-fold cross validation (see Table 2). The reliable results demonstrate the usefulness of MDA for LMI prediction, and the effectiveness of the proposed model to integrate different types of biological networks for prediction.

Table 2 Prediction performance on LMI dataset using MVMTMDA in k-fold cross validation

CV method	2-fold CV	5-fold CV	10-fold CV
Average AUC	0.8747+/-0.018	0.9014+/-0.012	0.9037+/-0.011

In this subsection, we compare the proposed MVMTMDA with other methods that were previously proposed for predicting MDA and LMI. There are an increasing number of computational tools proposed for predicting potential microRNAs involved in different diseases. We here select four methods for performance comparison, all of which are recently published in 2018.

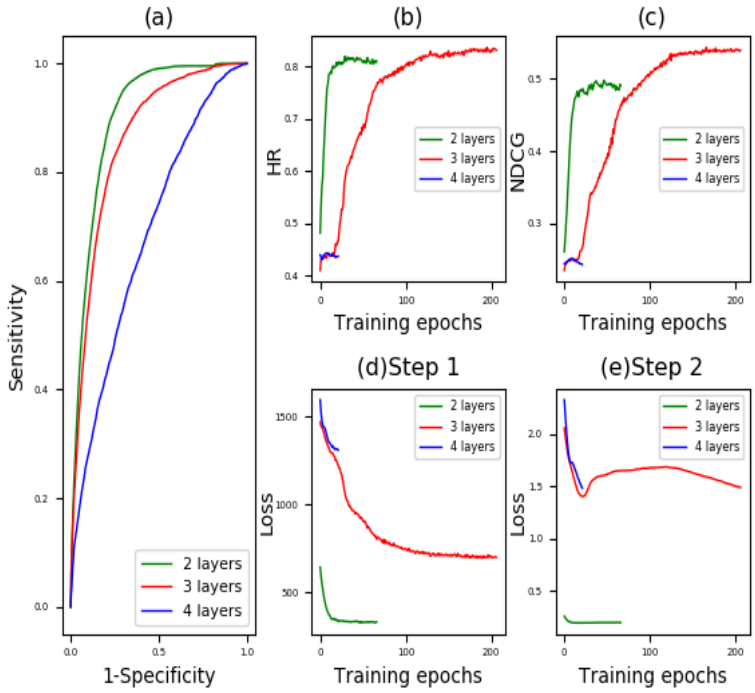

Figure 3 Performance yielded MVMTMDA in LMI prediction: (a)ROC curves yielded by MVMTMDA with 2, 3 and 4 layers; (b) Hit ratio yielded by MVMTMDA with increasing training epochs; (c) NDCG yielded by MVMTMDA with increasing training epochs; (d) the training loss

In these works, there are two kinds of data used to compute the microRNA similarity. One is the microRNA functional similarity yielded by MISIM [18], which hasn't been updated for several years. In addition, Wang's microRNA similarity was calculated based on an MDA dataset collected in 2010 such that it is inappropriate for MDA prediction. The other one is microRNA sequence similarity. However, the relation between microRNA functional similarity on pathology and microRNA sequence similarity is still unknown. To the best of our knowledge, the proposed MVMTMDA is the first one to use LMI to predict MDA. For the sake of fairness, we also introduced LMI into the compared prediction models.

Apart from the prediction tools for RNA target that are based on sequence matching, existing network-based prediction models for LMI is limited. For the performance

comparison about LMI prediction, we compare MVMTMDA with the model of EPLMI and three other baseline methods (i.e. Katz measure, basic latent factor model and neighbor-based collaborative filtering).

Different from the MVMTMDA model adopting end-to-end learning, these comparison methods need a microRNA similarity matrix as input. To execute the comparison methods on our datasets, we first construct a microRNA similarity matrix *MS* using Pearson correlation coefficient (PCC) as follow.

$$MS(i,\mathrm{j}) = \frac{\sum_{k=1}^{N}(S_{ki}-\bar{S}_{\cdot i})(S_{kj}-\bar{S}_{\cdot j})}{\sqrt{\sum_{k=1}^{N}(S_{ki}-\bar{S}_{\cdot i})^2 \sum_{i=1}^{N}(S_{kj}-\bar{S}_{\cdot j})^2}} \tag{17}$$

where *S* denotes the matrix of side information. That is, it denotes the adjacent matrix of LMI when predicting MDAs and, on the other hand, the adjacent matrix of MDA when predicting LMIs. As a result, the compared methods yielded AUCs arranging from 0.6233 to 0.8192 in MDA prediction and AUCs arranging from 0.7301 to 0.8737 in LMI prediction, both of which are significantly low than those yielded by MVMTMDA (see Table 3).

The reasons for the superior performance of the proposed model may lie in two aspects. One is that MVMTMDA adopts deep neural network structure which can automatically learn the complex relation between microRNA from MDA and LMI network in an end-to-end manner. The other one is that the proposed model considers the incompleteness of the side information and adopts multi-task learning to fill the missing values of it.

Table 3 Performance comparison on the prediction of MDA and LMI in 5-fold cross validation.

Prediction task	Method	Average AUC
Prediction of microRNA-disease associations	IMCMDA [40]	0.6233+/-0.032
	MDHGI [109]	0.6932+/-0.027
	Zeng's work [110]	0.7883+/-0.012
	MDA-SKF [111]	0.8192+/-0.010
	The proposed method	0.8512+/-0.012
Prediction of lncRNA-microRNA interactions	Neighbor-based CF [112]	0.7301+/-0.026
	LFM CF [113]	0.7692+/-0.025
	EPLMI [12]	0.8126+/-0.012
	Katz [114]	0.8737+/-0.008
	The proposed method	0.9014+/-0.012

3.3.3. Impact of side information on MVMTMDA

As mentioned in section 4.1 and 4.2, MVMTMDA predicts MDAs using the network of LMI as side information and can also predict LMIs with MDA network as side information. In this subsection, we evaluate the usefulness of the introduction of the side information. Specifically, for performance comparison, the second step of optimization (Equation 15) is discarded, such that the data of side information would be ignored when training the model. As shown in Table 4, without using the side information, the prediction performance of the proposed model significantly declines in the 2-fold and 5-fold cross validation. The comparison results demonstrate the ability of MVMTMDA to integrate multiple graph data, and also confirms our assumption that the information of LMI and MDA is closely related and mutually beneficial for the prediction task of each other.

Table 4 Results of 2-fold and 5-fold cross validation yielded the proposed model with and without side information

Prediction	Cross validation	MVMTMDA with side information	MVMTMDA without side information

MDA prediction	2-fold CV	AUC: 0.8410; HR: 0.7196; NDCG: 0.4429	AUC: 0.8306; HR: 0.7224; NDCG:0.4507
	5-fold CV	AUC: 0.8512; HR:0.7553; NDCG: 0.4895	AUC: 0.8423; HR: 0.7442; NDCG: 0.4705
LMI prediction	2-fold CV	AUC: 0.8747 HR:0.731; NDCG: 0.4119	AUC: 0.8316; HR:0.6217; NDCG: 0.3445
	5-fold CV	AUC: 0.9014; HR:0.8506; NDCG: 0.5542	AUC: 0.8697; HR:0.8291; NDCG: 0.5470

3.3.4. Sensitivity to Hyper-Parameters

Depth of layers in networks

The number of layers in neural networks is critical for the performance of deep learning-based models. In this work, we simply set the layer numbers and the layer sizes of neural networks α, β and γ the same. We set the number layers as 2, 3 and 4 for testing. Table 5 shows the prediction performance yielded by MVMTMDA with different layers in 5-fold cross validation. Figure 2 and 3 show the corresponding curves for prediction performance and optimization. The results show that the proposed model was optimized with layer number set as 2. We therefore use such structure for MVMTMDA in the experiments of this paper.

Table 5 Prediction performance using MVMTMDA with 2, 3 and 4 layers in 5-fold cross validation.

Prediction task	Depth of Layers in Networks		
	2 layers	3 layers	4 layers
MDA prediction	0.8512+/-0.012	0.781+/-0.011	0.8384+/-0.015
LMI prediction	0.9014+/-0.012	0.8602+/-0.015	0.7647+/-0.022

Negative Sampling Ratio

In this work, the samples from the datasets of MDA and LMI we collected are all positive such that the prediction task is a semi-supervised learning problem in which unlabeled samples are important to be considered. To training the model, we need to sample negative instances from unlabeled data to construct the set of X^- in Equation 14 and 15. In this experiment, we apply different negative sampling ratios (i.e. 1, 3 and 5) to observe the performance variance with regards to the prediction on MDA and LMI. As shown in Table 6, MVMTMDA yielded the best prediction performance with negative sampling ratio set as 1 and 5 in the prediction of MDA and LMI, respectively. The prediction performance is generally stable with different negative sampling ratios.

Table 6 Prediction performance using MVMTMDA with 2, 3 and 4 layers in 5-fold cross validation.

Prediction task	Negative sampling ratio		
	1-neg	3-neg	5-neg
MDA prediction	0.8512+/-0.012	0.8506+/-0.011	0.8437+/-0.015
LMI prediction	0.9014+/-0.012	0.8922+/-0.015	0.9109+/-0.012

3.3.5. Functional clustering of microRNAs based on multi-view embedding features

In recent years, the similarity measurement for function of microRNAs has attracting increasing attention due to its significance in the domain of noncoding RNA research[115, 116]. In this section, we propose a new type of functional similarity measure for microRNAs based on MDA and LMI.

As the connection joint of the networks of MDA and LMI, in this work, microRNA is considered to have two views to represent its biological functions. Motivated by this,

the proposed MVMTMDA learns graph embedding features for each type of microRNA, comprehensively considering their relationship with diseases and target lncRNAs. The microRNA features learned from the proposed model can thus imply the functional similarity among microRNAs. In this section, we implement K-means clustering in the feature space of the microRNA graph embedding learned by the MVMTMDA model.

Specifically, we first use all data in MDA dataset as training set to train the of model of MVMTMDA until results converged. Secondly, we applied principal component analysis (PCA) on microRNA features. Based on the first three dimensions in PCA, the clustering algorithm of k-means was implemented. We set the number of clusters as 6 and the corresponding scatter diagram is shown in Figure 4. In addition, we calculate the PCC of microRNA features as the function similarity score. It is anticipated that the microRNA-disease pair with high predicted scores will be confirmed by biological experiment in the future.

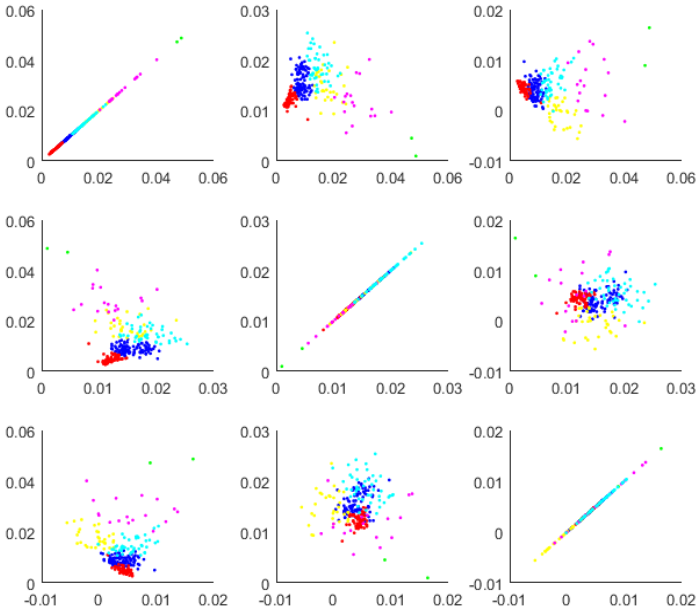

Figure 4 Scatter diagram of functional clustering for 268 types of microRNAs

3.4. Summary

The identification of MDAs is of great significance in microRNA therapeutics. Current computational methods for predicting MDAs haven't considered the co-regulation between lncRNA and microRNA, which is becoming known to be very important for their function mechanisms. In this work, we propose a multi-view multi-task model composed of three deep neural networks to fill this gap. Considering the networks of MDA and LMI are two different views collaboratively implying the biological function of microRNAs, we apply a multi-view learning method to extract embedding features for microRNA from two different graphs. In addition, we combine the prediction of MDA and LMI, which are closely related as they both belong to parts of aberrant ceRNA regulation on diseases. A number of experiments were implemented on the real datasets that we collected and extensive analysis is also made on the predicted results.

The experimental results demonstrate of the feasibility and effectiveness of the proposed model to predict MDA on a large scale.

The main contribution of our work is fourfold. Firstly, the propose model is the first one to consider the interaction between lncRNA and microRNA for large-scale prediction of MDA. LMI is ideal data to uncover the association between microRNA and disease due to their meaning and data type. Secondly, we consider the incompleteness of the side information and use a multi-task learning method to synchronously predict MDAs and LMIs. Thirdly, the proposed model enables an end-to-end prediction for MDA. Any type of graph data associated with microRNA (e.g. microRNA-gene interaction and microRNA-protein interaction) can be flexibly and directly used as inputs to improve the prediction, which is important because the amount of microRNA data is increasing rapidly. Fourthly, different from similarity-based model, the proposed model can automatically extract features from the raw data, providing a new type of data source for measuring microRNA functional similarity.

4. GCMDR - A GRAPH CONVOLUTION-BASED ALGORITHM FOR PREDICTING ASSOCIATIONS BETWEEN MICRORNA AND DRUG RESISTANCE

4.1. Background

The field of pharmacogenomics has evolved from the observing of variable drug responses to the development of modern molecular medicine [117]. The recent focus on finding associations between individual genomic and transcriptomic features and the efficacy and toxicity of a drug, and this is useful in facilitating the development of personalized treatment strategies. In recent years, the cost of drug development keeps going up and the main reason for the relatively poor productivity in R&D in the pharmaceutical industry lies in the difficulty in drug-target selection. Among approved drugs, more than 80% of them are developed to target only at the proteins of enzymes and receptors and greater than 99% of them target at some specific proteins [4]. Thus far, although human genome has been found to encode up to 25,000 genes, current drugs can only target at about 600 types of disease-modifying proteins [118, 119]. In other words, a considerable amount of proteins are “undruggable”. Hence, the focus in target selection has now shifted to other macromolecules including noncoding RNAs (ncRNAs). Specially, microRNAs which is a type of ncRNAs are identified as potential high-value targets for therapy due to its involvement in gene regulation, and as a result, microRNA pharmacogenomics is starting to become the new frontier for personalized

medicine [33].

Despite considerable advances in targeted-therapy techniques and the knowledge accumulated about molecular carcinogenesis of human diseases, there is still a large gap between the identification of microRNA interacting pathway and practical therapeutic use. The evidence cumulated has demonstrated that the variations in microRNA profiling of patients can be a major cause of individual differences in drug sensitivity or resistance [2]. Overexpressed microRNAs can downregulate genes with protein products necessary for drug efficacy. Conversely, insufficient microRNA expression can also upregulate genes with protein products inhibiting drug function [120]. As both increased and decreased microRNA expression levels can cause diseases, accordingly, the microRNA-targeted therapeutics agents can be divided into microRNA mimics and inhibitors, which respectively aim to induce gene silencing and selective upregulation of proteins [121]. As mentioned above, not every protein can be targeted or modulated by drugs. Therefore, precise manipulation of expression levels of microRNAs associated with these “undruggable” proteins harbors great theranostic implications [4].

The primary challenges faced by the current microRNA-target therapeutics are twofold: the successful delivery of therapeutic agent to the target tissues and the safety evaluation of potential drug response [122]. The problem of poor cell-permeability and pharmacokinetics in the first challenge can be partially solved by following appropriate parameters for molecule designing, which can be summarized by Lipinski’s Rule of

Five [118]. The second challenge here is the need for thorough understanding about the impact of microRNA expression on drug response. However, for most of drugs, little is known about their drug resistance in relation to different microRNA profiling. It is a multifactorial phenomenon involving several inducements such as decreased uptake of water-soluble drugs, increased repair of DNA damage, increased energy-dependent efflux of drugs and altered metabolism of drugs [123, 124]. An increasing number of studies have investigated differences in microRNA expression levels affecting the drug response. Recently, a comprehensive database called ncDR has been built to record microRNA-drug resistance associations, providing data resource for further computational analysis [7].

Although an increasing number of microRNA types have been identified to be associated with drug resistance spanning almost all classes of drugs, available knowledge on associations between microRNA and drug resistance is still far from sufficient to meet the requirement for guiding microRNA-targeted drug development. Computational methods could facilitate this discovery process, but little work has been done in this important direction. In this paper, we propose a novel computational method for inferring drug resistance-associated microRNAs, expecting to boost the efficiency and pace of microRNA-targeted drug discovery and development. To the best of our knowledge, this is the first of its kind.

By using data of known microRNA-drug resistance associations and various types of intrinsic features of microRNAs and drugs, we develop a deep learning-based

prediction model called GCMDR (Graph Convolution for association between MicroRNA and Drug Resistance) to predict new drug resistance-associated microRNAs. The basis assumption behind the development of the model is that different types of microRNAs induce drug resistance by following same latent mechanisms that drug structure and microRNA expression level are involved. With an orchestrated transcriptional regulation pattern, microRNAs of similar biological functions may suppress common expression products (such as membrane transporter proteins involving drug efflux and uptake) and can further affect responses to the same multiple therapeutic agents. The suppressed expression products may also be associated with other mechanisms including decreased drug cytotoxicity by detoxification or inactivation, alternations of regulation of cell cycle and checkpoints, resistance to apoptosis, enhanced DNA damage repair, which may modify disease response to therapeutic agents [8].

The concept that one microRNA regulates one gene that affects response to one drug is the basis of some experimental models, and this could confine targeted therapeutics to the 'one-drug-one-target' paradigm making them susceptible to resistance in due course [34]. GCMDR enables a more comprehensive drug-resistance analysis by taking the full scope of microRNAs into considerations. Specifically, given any drug with their structure information, GCMDR can effectively compute the putative level of drug resistance associated with all microRNAs, laying out the underlying rationale for the use of microRNA-targeted drugs to achieve an orchestrated broad gene silencing.

4.2. GCMDR in details

4.2.1. Challenges in predicting microRNA-drug resistance association

In order to predict associations between microRNAs and drug resistance, we need to tackle several challenges. First, the number of microRNA-drug resistance associations that is discovered and cataloged in the ncDR database is relatively small. In order to make use of the domain information related to microRNA-based drug response, we introduce various types of in-vivo/in-silico features for GCMDR to better perform its tasks. These features include microRNA expression profiles, drug PubChem substructure fingerprints, microRNA GO-based and disease-based functional features.

Second, the dimensions of the intrinsic features are so high that it is hard to be used directly for measuring how microRNAs or drugs are similar, and therefore traditional neighbor-based methods may not be able to effectively capture meaningful information from intrinsic feature inputs. To address this issue, grounded on the spectral graph therapy, we propose an end-to-end learning method based on a graph convolution operator, which automatically extract features from the raw data of multiple attributes of drug/microRNA and the microRNA-drug associations.

Finally, as microRNAs influence drug efficacy by regulating the expression of genes, the microRNA-gene and gene-drug associations are important for predicting the microRNA-drug resistance associations. However, these kinds of information are unavailable for most known microRNA-drug resistance associations. Considering this

structure of associations would be helpful for inferring whether a type of microRNA and a type of drug assistance are associated. Accordingly, we combine the techniques of auto-encoder and latent factor model, allowing GCMDR able to learn a hidden factor layer topologically similar with the real "coding gene layer".

4.2.2. Data collection

The dataset we used in this work was collected from the April 2018 version of the ncDR database which is publicly released at http://www.jianglab.cn/ncDR [7]. This dataset contains 5,864 relationships between drug compounds and ncRNAs which include 877 microRNAs and 162 lncRNAs obtained through manual curation from about 900 literatures. For the purpose of our work, we focus only on association relationships involving microRNAs and obtained 3,338 microRNA-drug resistance associations after removing 1,859 redundant ones. The resulting dataset obtained contain 754 different types of microRNAs and 106 different types of drug compounds.

In order for us to take into account the structure similarity between drugs so as to improve the prediction performance of GCMDR, we obtained the substructure fingerprints of the 106 drug compounds from the PubChem database (https://pubchem.ncbi.nlm.nih.gov) [125] for our work. We transformed the 64base-encoded PubChem substructure fingerprints into binary features represented in 920 bits.

To describe the properties of microRNAs for predicting their associated drug response, three types of information from various databases were obtained. However, not all 754

microRNAs in our collected dataset can be found in these databases. As an unavoidable step to build a model using graph convolution, for those microRNAs whose information is not available, we estimated the missing values as the average of those others whose values are known. To the common problem of data missing, we can choose from a number of different methods. We bring up listwise deletion, mean imputation, hotdecking, regression imputation, multiple imputation as examples. Of these different methods, mean imputation is the most popular method as it is simple and least computationally expensive and is widely used when handling large data sets [126]. Hence, we chose it to estimate missing values.

The first of such information is the microRNA expression profile collected from the microRNA.org database. It has 172 dimensions representing the expression levels of a single type of microRNAs in 172 different human tissues and cell lines. Each value of expression profile represents the number of cloned mature microRNAs that were sequenced in 172 small RNA libraries and reported as normalized clone counts. Out of the 754 microRNAs in the collected dataset, we managed to find 540 in the miroRNA.org database (http://www.microrna.org/microrna/home.do) [11].

For performance assessment, we have also collected two other types of microRNA features, both of which are functional features obtained by previously proposed computational methods. One of them was obtained from the work described in [127] which reports on a 2589×2589 microRNA functional similarity matrix computed based on the Gene Ontology (GO) terms. From it, we were able to extract the GO-based

functional features of 497 of the 754 microRNAs. Each GO-based microRNA feature has 2589 dimensions and the similarity scores of a microRNA to other 2588 microRNAs with respect to their gene regulation functions were obtained.

Another type of microRNA functional feature considered in this work are the microRNA-disease associations. There have been evidence that microRNAs corporately function in the mechanism of human diseases[128, 129]. It is for this reason that some attempts have been made to elucidate microRNA functional similarity by using domain knowledge related to diseases, such as disease ontology (DO). A web-based bioinformatics toolkit called DincRNA (http://bio-annotation.cn:18080/DincRNAClient) [116] has recently been proposed to provide such kind of microRNA functional similarity calculations. From there we obtained eight types of DO-based microRNA features computed using eight different algorithms. We found matches of the ID of 332 microRNAs in our dataset of 754 with that in DincRNA. Each DO-based microRNA feature has 556 dimensions, representing the similarity scores of one microRNA to other 555 microRNAs according to their involvement in disease mechanism.

4.2.3. The concept of graph convolution

By representing known associations between microRNA and drug resistance as a bipartite graph, the prediction problem we consider in this paper can be defined as a task related to semi-supervised link prediction on such a graph.

Let us assume that we are given a bipartite graph $\boldsymbol{G} = (\boldsymbol{v}, \boldsymbol{\varepsilon})$ with $\boldsymbol{v} = (\boldsymbol{v_m}, \boldsymbol{v_d})$ representing n_m microRNA nodes and n_d drug nodes, which have numerical node features $X_m = \left[x_m^1, x_m^2, \ldots, x_m^{n_m}\right]^T \in \mathbb{R}^{n_m \times c_m}$ and $X_d = \left[x_d^1, x_d^2, \ldots, x_d^{n_d}\right]^T \in \mathbb{R}^{n_d \times c_d}$ respectively. Suppose that the labels of some links, $\boldsymbol{\varepsilon}$ in $\boldsymbol{G}$ are given, the goal is to predict if there is any potential link between any microRNA-drug pair that have not yet previously been established. As the dimensions of feature vectors of microRNA and drug nodes can be as high as 2589 and 920 respectively, traditional similarity measure, like the Euclidean distance, would not be very ineffective as the contribution of important feature values to the similarity measure would be ignored and this is similar to the problem of image retrieval in which the data are matrices of pixels. Hence, the problem of how best to effectively utilize both graph topology and the attribute information of the nodes need to be addressed. Towards this goal, we propose to use a spectral graph convolution

Current approaches to designing localized convolutional filters on graphs can be classified roughly into two categories: the spatial and the spectral approach. The former provides filter localization by using local information of neighboring vertexes such as that used in Niepert's work [83]. The problem with such approach of matching local neighbors are pointed out in [9]. As opposed to the spatial approach, the spectral approach is mainly designed based on the spectrum of the graph Laplacian. This approach provides a well-defined localization operator on graphs using a Kronecker delta implemented in the spectral domain.

There have recently been some attempts to use deep learning techniques to graph-based

data analytics. In [85], a graph convolutional neural network (GCNN) has been proposed. Graph convolution is defined on graph as the multiplication of an input signal with a filter $\boldsymbol{g_\theta}$ in the Fourier domain. Given an adjacent matrix $\boldsymbol{A}$ with its Laplacian $\boldsymbol{L} := \boldsymbol{D} - \boldsymbol{A}$, and attributes of each node on graph (say $\boldsymbol{s}$), spectral graph convolution tries to decompose $\boldsymbol{s}$ on the spectral domain and then design and apply a spectral filter function $\boldsymbol{g_\theta}$ on the spectral components. Suppose that $\boldsymbol{L}$ can be decomposed by $\boldsymbol{L} = \boldsymbol{U\Lambda U^T}$, where $\boldsymbol{\Lambda}$ is the diagonal matrix of eigenvalues and $\boldsymbol{U}$ is eigenvector matrix. Here, $\boldsymbol{U^T s}$ could be considered as a graph Fourier transform of $\boldsymbol{s}$. $\boldsymbol{g_\theta}$ can be defined as $\boldsymbol{g_\theta \star s = U g_\theta U^T s}$. To circumvent this problem of computationally expensive eigendecomposition of $\boldsymbol{L}$, Defferrard et al. [85] approximate the spectral filter by using a truncated expansion in terms of Chebyshev polynomials $\boldsymbol{T_k(s)}$ up to K^{th} order:

$$g_\theta \star s \approx \sum_{k=0}^{K} \theta'_k T_k(L_N) s \tag{18}$$

where $\boldsymbol{T_k}$ is the Chebyshev polynomials and $\boldsymbol{\theta'}$ is a vector of Chebyshev coefficients. Kipf and Welling [86] further simplified this definition by limiting K=1 and approximating the largest eigenvalue of L by 2. The convolution operator comes to be:

$$g_\theta \star s = \theta(I + D^{-\frac{1}{2}} A D^{-\frac{1}{2}}) s \tag{19}$$

By introducing the renormalization tricks: $I_N + D^{-\frac{1}{2}} A D^{-\frac{1}{2}} \rightarrow \widetilde{D}^{-\frac{1}{2}} \tilde{A} \widetilde{D}^{-\frac{1}{2}}$ with $\tilde{A} = A + I_N$ and $\widetilde{D}_{ii} = \sum_j \tilde{A}_{ij}$, formula (2) can be simplified as:

$$g_\theta \star s = \theta \widetilde{D}^{-\frac{1}{2}} \tilde{A} \widetilde{D}^{-\frac{1}{2}} s \quad (20)$$

We adopted this simplified definition for graph convolution in this work.

Increasing attempts have recently been made to reveal the intrinsic and advantages of spectral graph convolution. For example, Li et al. have discovered that spectral graph convolution is a special form of symmetric Laplacian smoothing, which is the key reason why it works [130]. Atwood and Towsley claim that graph convolution can actually be explained as a graph diffusion kernel [131]. The first attempt to use graph convolution to develop bioinformatics tools was made by Duvenaud et al. [132]. They used it to learn fingerprint representation of chemical compounds which can be represented as graphs. It is anticipated that graph convolution will be extensively used in the field of bioinformatics.

4.2.4. Model structure of GCMDR

As mentioned above, the prediction of associations between microRNAs and drug resistance can be considered as a semi-supervised link prediction problem. Current GCNN-based methods have been used mainly to tackle node classification problem on homogeneous network and are thus not applicable to our problem involving prediction of associations between microRNA and drug resistance. To make use of GCNN, we have to extent the current graph convolution idea to allow it to solve link prediction problem defined on heterogeneous, bipartite, attributed networks. Towards this goal, we propose an GCMDR algorithm.

Given a graph representing known associations between microRNA and drug resistance with the corresponding adjacent matrix $\boldsymbol{M}$ of shape $n_d \times n_m$, the goal of GCMDR is to learn embedding features for microRNAs and drugs $\boldsymbol{F}$ by building a graph convolution-based encoder $[\boldsymbol{F_d}, \boldsymbol{F_m}] = \boldsymbol{f_{en}}(\boldsymbol{v}, \boldsymbol{\varepsilon}, \boldsymbol{X_d}, \boldsymbol{X_m})$ to predict new links by building a decoder $\boldsymbol{M'} = \boldsymbol{f_{de}}(\boldsymbol{F_d}, \boldsymbol{F_m})$ where $\boldsymbol{F_d}$ and $\boldsymbol{F_m}$ are the feature matrices for drugs and microRNA with shapes of $n_d \times L$ and $n_m \times L$, respectively (see figure 5). GCMDR is the first attempt to combine the techniques of graph convolution and auto-encoder.

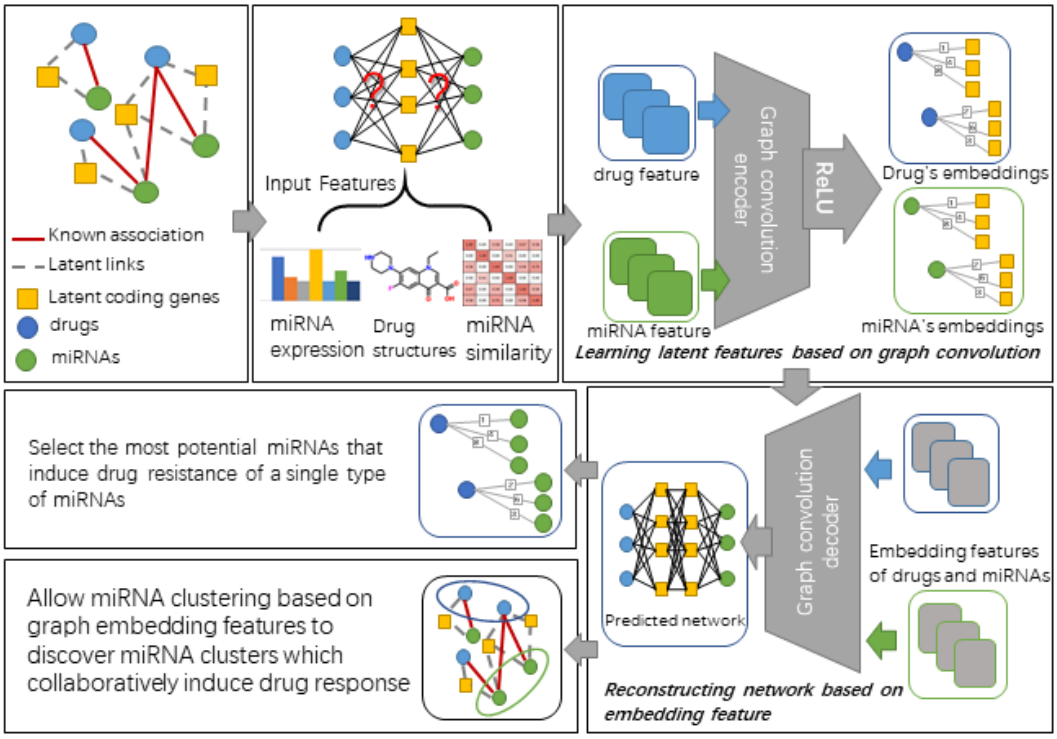

Figure 5 Flowchart of the proposed GCMDR model

To build an encoder, we propose to introduce graph convolution into the latent factor model in accordance with the nature of "microRNA-coding gene-drug resistance" associations. Specifically, an adjacent matrix $\boldsymbol{A}$ and feature matrix $\boldsymbol{X}$ is reconstructed based on $\boldsymbol{M}$, $\boldsymbol{X_d}$ and $\boldsymbol{X_m}$ as follows:

$$A = \begin{bmatrix} 0 & M \\ M^T & 0 \end{bmatrix} \tag{21}$$

$$X = \begin{bmatrix} X_d & 0 \\ 0 & X_m \end{bmatrix} \tag{22}$$

GCMDR then normalizes rows of matrix $\boldsymbol{X}$: $\mathrm{X}_{rw} := D^{-1}X$ with $D = \sum_j X_{ij}$. Here, $\boldsymbol{X_{rw}}$ represents the matrix of input signal with a shape of $(n_d+n_m)\times(c_d+c_m)$. According to (3), the graph convolution of the $\boldsymbol{X_{rw}}$ matrix of a graph with adjacent matrix $\boldsymbol{A}$ can be defined as:

$$F = \mathrm{X}_{rw}\left(I + D^{-\frac{1}{2}}AD^{-\frac{1}{2}}\right)W_e \tag{23}$$

where the trainable weight matrix $\boldsymbol{W_e} \in (c_d+c_m)\times L$ is the Fourier coefficient matrix. It transforms matrix $\mathrm{X}_{rw}\left(I + D^{-\frac{1}{2}}AD^{-\frac{1}{2}}\right)$ into a hidden matrix $\boldsymbol{F}$ which describes the hidden association between drug or microRNA nodes and the latent factors (coding gene layer). The n_e denotes the number of latent factors and is set manually. By introducing the activation function of ReLU and bias matrix $\boldsymbol{B}$ to the hidden matrix $\boldsymbol{F}$, the embedding feature matrices of drug nodes and microRNA nodes, $\boldsymbol{F_d} \in n_d\times L$ and $\boldsymbol{F_m} \in n_m\times L$ could be obtained as follow:

$$\begin{bmatrix} F_l \\ F_m \end{bmatrix} = \mathrm{ReLU}(F + B) \tag{24}$$

To reconstruct the adjacency matrix for drug-microRNA associations, a decoder $\boldsymbol{M'} = \boldsymbol{f_{de}}(\boldsymbol{F_d}, \boldsymbol{F_m})$ is built as follow:

$$M' = F_l W_d {F_m}^T \tag{25}$$

where $\boldsymbol{W_d} \in L \times L$ is a weight matrix describes the similarity between latent factors in hidden "coding gene" layer. In this work, we randomly initialize all trainable matrices (i.e. $\boldsymbol{W_e}$, $\boldsymbol{W_d}$ and $\boldsymbol{B}$) using the initialization approach described as in [133].

In addition, we also implement negative sampling for model training. Specifically, in each training epoch, unlabeled drug-microRNA pairs are chosen randomly to be negative samples for training. Given these training data, GCMDR attempts to minimize the following loss function:

$$\mathcal{L} = \sqrt{\frac{\sum_{ij;\Omega_{p,ij}=1 or \Omega_{n,ij}=1}(M_{ij}{}' - M_{ij})^2}{\sum_{ij}(\Omega_{p,ij}+\Omega_{n,ij})}} + \frac{1}{2}\|W_e\|^2 + \frac{1}{2}\|W_d\|^2 + \frac{1}{2}\|B\|^2 \qquad (26)$$

where the matrices $\Omega_p \in \{0,1\}^{n_d \times n_m}$ and $\Omega_n \in \{0,1\}^{N_l \times N_m}$ serve as the masks for positive samples and the negative samples from random sampling, respectively. Also, the first term in equation (26) aims to minimize prediction errors and the rest of the other terms define constraints on the weight matrices in encoder and decoder, respectively. As negative sampling was implemented for training, in each epoch the Ω_n would be randomly generated in which the number of "1" can be fixed to be 10 times the number of positive samples. Hence, optimization in GCMDR is performed over the positive samples if we set this percentage to be 0 or over the positive samples and partial negative samples if we set this percentage to be larger than 0.

4.3. Experiment and analysis

To evaluate the performance of GCMDR, k-fold cross validations are used. Specifically,

in each round of k-fold cross validation, a fixed number of known associations between microRNA and drug resistance take turns to be used as testing samples (association relationship assumed to be unknown) so that prediction scores can be used using GCMDR. These prediction scores can indicate how possible the prediction of relationship is positive. In our performance evaluation, testing samples along with all unlabeled microRNA-drug pairs are considered as candidate samples. We consider the results as demonstrating good performance if the testing samples are ranked high among all candidate samples in terms of their prediction scores.

To generate training and testing data sets, all known microRNA-drug resistance associations are randomly divided into *k* subsets of roughly the same sizes. Prediction experiments were repeated *k* times using different data sets in which each subset takes turn to be used as training samples and the remaining subsets used as testing samples. To decide if a testing sample is positive, its prediction score is compared with other candidate samples. If it is ranked higher than the given threshold, the existence of association relationship would be considered highly possible.

By setting different thresholds, we computed corresponding true positive (TPRs, sensitivity) and false positive rates (FPRs, 1-specificity) where sensitivity and specificity denote the percentages of testing samples with respectively higher and lower ranks than the given thresholds. The corresponding receiver operating characteristic (ROC) curves have been obtained by plotting TPR versus FPR. To measure prediction performance with a ROC curve, we computed its AUC with ranges from 0.5 to 1, where

0.5 denotes a purely random prediction and 1 denotes perfect performance. In this work, 2-fold, 5-fold, 10-fold cross validations are implemented.

4.3.1. Similarity-based methods compared with GCMDR

Although predicting microRNA-drug resistance is a new problem that has not been tackled by researchers in bioinformatics before. However, one may make use of different prediction methods developed for graph-based data e.g. approaches developed to find ncRNA-protein interactions [134], microbe-disease association [135] and lncRNA-microRNA interactions [136]. However, almost all these methods requires that similarity matrices be computed by predefining specific similarity measure, which is crucial for the performance. However, the dimensions of drug feature and microRNA feature we collected in this work are too huge so that these preprocessing would be harmful to the prediction performance. We therefore adopt graph convolution to build an end-to-end prediction model to circumvent this problem.

To further evaluate the prediction performance of GCMDR, we compared its performance against six other similarity-based methods. The results obtained with these methods were used as baseline results and these methods include three collaborative filtering (CF) methods (i.e. drug-based CF, ncRNA-based CF and neighbor-based CF) [137], one matrix factorization-based method (i.e. SVD with reserved ranks of 10) [138] and two graph diffusion-based methods (i.e. Katz and EPLMI) [135, 136]. As data preprocessing, similarity matrices of microRNAs and drugs were constructed by using Pearson correlation coefficient (PCC) as follow:

$$S_*(a,b) = \frac{\sum_{i=1}^{N}(f_{ai}-\overline{f_a})(f_{bi}-\overline{f_b})}{\sqrt{\sum_{i=1}^{N}(f_{ai}-\overline{f_a})^2 \sum_{i=1}^{N}(f_{bi}-\overline{f_b})^2}} \quad (27)$$

where f_a and f_b are two features of two elements (a and b) from a microRNA-microRNA or drug-drug pair. Based on the PCC-based similarity matrices for microRNAs and drugs (i.e. $\boldsymbol{S_{microRNA}}$ and $\boldsymbol{S_{drug}}$), the predicted score matrix yielded by drug-based CF can be defined as:

$$M'_{drug}(d_i, m_j) = \frac{\sum_{k=1}^{n_d} S_{drug}(d_i,d_k)\cdot M_{k,j}}{n_d} \quad (28)$$

Similarly, the score matrix predicted by microRNA-based CF can be defined as:

$$M'_{mirna}(d_i, m_j) = \frac{\sum_{k=1}^{nm} S_{mirna}(m_j,m_k)\cdot M_{i,k}}{n_m} \quad (29)$$

Neighbor-based CF considers both collaborative effects of drugs and microRNAs and can be defined as:

$$M'_{neighor} = \frac{M'_{drug}+M'_{mirna}}{2} \quad (30)$$

Katz metric is a classical method which is initially proposed to tackle social network problems. It is now extensively used to solve bioinformatics problems such as those involving prediction of microbes and ncRNAs associated with diseases [39, 41]. EPLMI is a two-way diffusion problem which is proposed to predict new lncRNA-microRNA interactions.

4.3.2. Performance evaluation for GCMDR

To evaluate the accuracy of the prediction model constructed by GCMDR, we used a real dataset involving experimentally-conformed microRNA-drug resistance associations. The prediction accuracy of the model was tested using 2-fold, 5-fold and 10-fold cross validation. The results presented in this section were obtained when GCMDR was used to construct the models using features of microRNA expression profile and drug PubChem substructures. As shown in figure6(a) and figure 6(b), for all three experiments, the training loss and the prediction error were found to be convergent when the training epoch reached about 150.

To avoid any bias caused by the random generation of training sets in the cross validation process, we repeated random sampling along with prediction for 20 times. The average values of AUCs using different cross validations were shown in Table 7. It is known that prediction accuracy increases with the size of the training data. As the training set for 10-fold CV is larger than that for 2-fold and 5-fold CV, the average AUC that the 10-fold CV yields is the highest and it is 0.9369 ± 0.0003.

These results demonstrate that GCMDR is a promising approach for inferring new microRNAs based on the resistance of specific drugs that it associates with. Considering that a full-scope analysis on microRNA-drug resistance is significant for drug development and that there is still little effort made in this important direction. They were obtained by GCMDR constructing models based on known microRNA-drug resistance associations, microRNA expression profiles and drug PubChem substructure

fingerprints as inputs. Candidate associations that are ranked high in the matrix are expected to eventually be confirmed to be real associations by laboratory experimentation in the near future.

Table 7 Prediction performance w.r.t. AUC using different kinds of cross validation

CV methods	2-fold CV	5-fold CV	10-fold CV
Average AUC	0.9301+/-0.0005	0.9359+/-0.0006	0.9369+/-0.0003

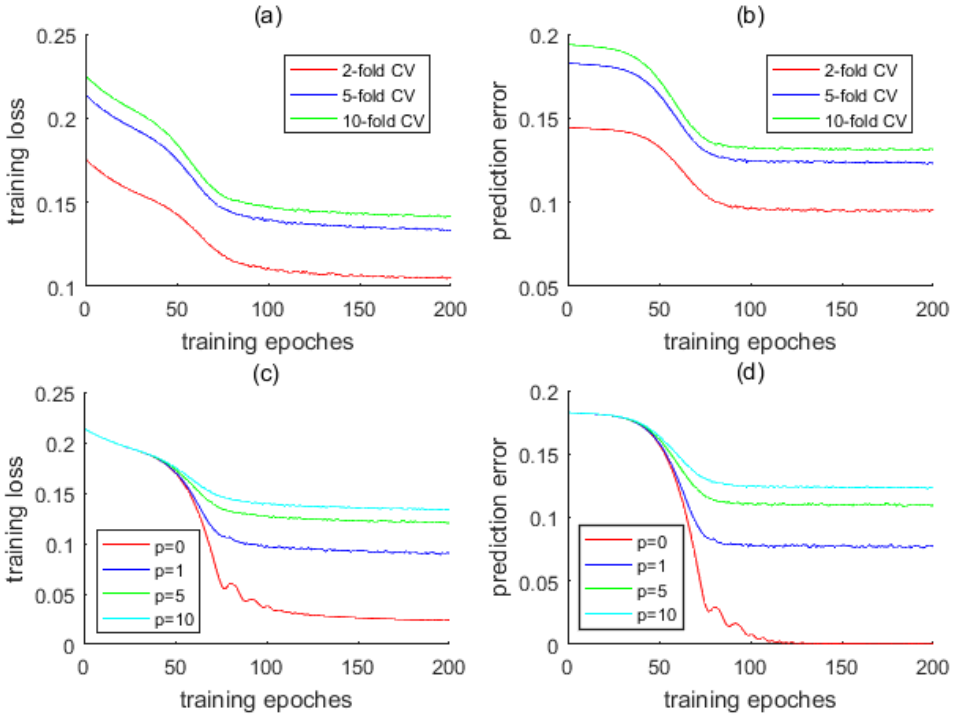

Figure 6 Training process w.r.t. training loss and training error. (a) and (b) show results of the training process corresponding to 2-fold; 5-fold and 10-fold cross validations and (c) and (d) show results of the training processes corresponding to different settings of negative sampling.

4.3.3. Comparison with microRNA features based on functional similarity

MicroRNAs are found to collaboratively deregulate gene expression and are, therefore, involved in the mechanism of diseases. Motivated by this, some efforts have started to be made to determine the similarity between microRNAs based on prior knowledge of gene ontology and disease MeSH ontology. Like the use of expression profiles, such

kind of microRNA similarity can be considered as a special feature of a microRNA with respect to their biological functions.

In this section, we explain how six different algorithms can be used to compute 11 different types of similarity measures between microRNAs. In computing each of these measures, drug substructures were also considered as feature inputs by GCMDR. As discussed in Section 4.1, 5-fold cross validation was repeated 20 times in our experiments so that we can compute an average AUC for performance evaluation. Table 2 shows the results of performance-comparison among these features. From the results, it can be noted that all ten types of DO-based microRNA features yielded similar performance with average AUCs slightly higher than 0.930. By introducing GO-based microRNA features, we show that this could lead to relatively better performance with average AUCs of 0.9353, which is much closer to that obtained with microRNA repression profiles. These results indicate that GO terms are more effective for describing microRNA properties w.r.t. drug resistance than DO terms. This may be due to the fact that coding genes have a more direct relationship with microRNAs than disease symptoms and therefore, the GO terms can be used to more effectively describe the functional properties of microRNAs.

Other than the above, an additional experiment was performed to evaluate the effectives of graph convolution in improving the prediction performance. Specifically, we replaced every entity in the feature matrix (i.e. $\boldsymbol{F}$ in equation 6) with '1' and kept the rest of the other steps the same for prediction. This was done to ensure that the graph

convolution operator played no effect in prediction. As expected, without any feature input, GCMDR model yielded average AUCs of 0.9257±0.0007, which is significantly lower than that obtained with the use of graph convolution with feature inputs (see Table 7 and figure 5(b)). These results demonstrate the ability of graph convolution to incorporate the information of intrinsic feature inputs and the graph topology.

Table 8 Performance comparison w.r.t. AUC in 5-fold CV among different types of microRNA feature input using GCMDR

Data type of microRNA features		Average AUC
No feature input		0.9257+/-0.0007
Expression profile		**0.9359+/-0.0006**
Gene ontology-based microRNA similarity [127]		0.9353+/-0.0005
Disease ontology-based microRNA similarity [116]	Lin_PAPM [139]	0.9277+/-0.0002
	Resnik_PAPM [140]	0.9307+/-0.0005
	Wang_PAPM [10]	0.9307+/-0.0004
	PSB_PAPM [141]	0.9306+/-0.0006
	SemFunSim_PAPM [142]	0.9305+/-0.0005
	Lin_PBPA [139]	0.9309+/-0.0005
	Resnik_PBPA [140]	0.9310+/-0.0006
	Wang_PBPA [10]	0.9313+/-0.0005
	PSB_PBPA [141]	0.9307+/-0.0006
	SemFunSim_PBPA [142]	0.9307+/-0.0006

4.3.4. Performance comparison between GCMDR and similarity-based methods

To evaluate the prediction performance of GCMDR, a series of comparison experiments were performed on the same dataset so that some baseline results can be obtained. As mentioned above, except for the SVD method which uses only the adjacency matrix $\boldsymbol{M}$, all different methods require that similarity matrices of microRNAs and drugs (i.e. $\boldsymbol{S}_{microRNA}$ and $\boldsymbol{S}_{drug}$) be computed. The 5-fold CV results for

the different methods are shown in Table 3 and figure 2(c). It is noted from these results that, among all seven different prediction methods, the models obtained with GCMDR yielded the highest prediction accuracy with the highest average AUCs of 0.9359±0.0006. These results indicate that GCMDR, with the benefit from the end-to-end computational structure, could be a reliable computational approach for the prediction of microRNA-drug resistance associations on a large scale.

Table 9 Performance comparison among different prediction methods w.r.t. AUC in 5-fold CV

Method	Average AUCs
GCMDR	**0.9359+/-0.0006**
EPLMI[136]	0.8971+/-0.0009
SVD-based MF	0.6007+/-0.0052
Katz metric	0.8471+/-0.0005
Drug-based CF	0.6490+/-0.0014
ncRNA-based CF	0.8103+/-0.0004
neighbor-based CF	0.8644+/-0.0009

4.3.5. Comparison among different numbers of latent factors

As GCMDR is built based on a latent factor model, the size of the latent layer would be crucial for its prediction performance. In this section, we evaluate the influence of the number of latent factors L on the prediction performance of GCMDR. Specifically, using the dataset described above, we compute the average AUCs based on a set of 5-fold CV experiments using features obtained from microRNA expression profile and drug structure. As Figure. 7 shows, the average AUC forms a unimodal distribution when L was set from 5 to 100, and L was at its optimal at 25. This number, 25, may reflect the number of functional clusters of genes in real associations between microRNA expression and drug resistance. In the rest of the other experiments, we

therefore set $L = 25$.

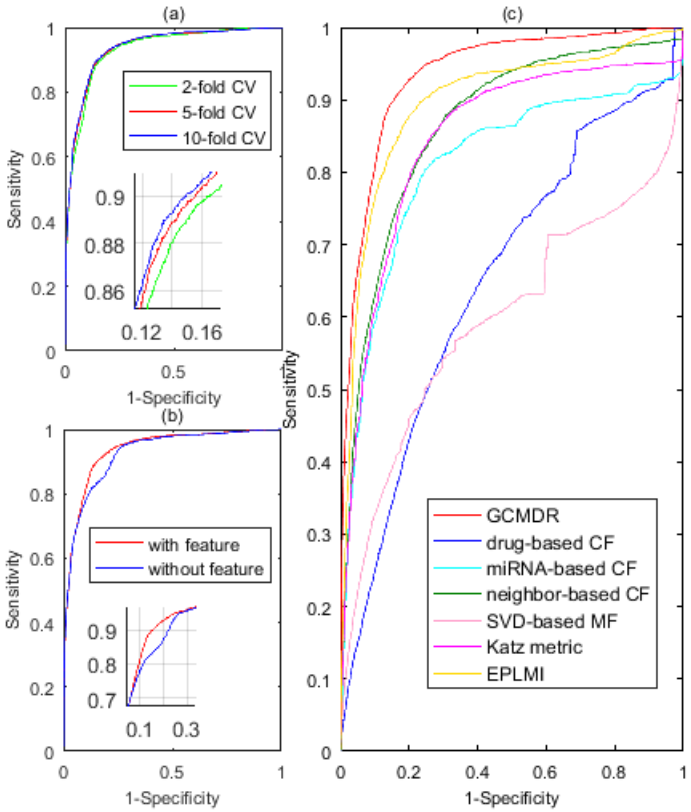

Figure 7 Prediction performance of GCMDR w.r.t. curves of ROC: (a)ROC curves yielded by GCMDR using 2-fold, 5-fold and 10-fold CV; (b) Difference of prediction performance using GCMDR with/without feature inputs; (c) Performance comparison of GCMDR with six types of similarity-based prediction methods.

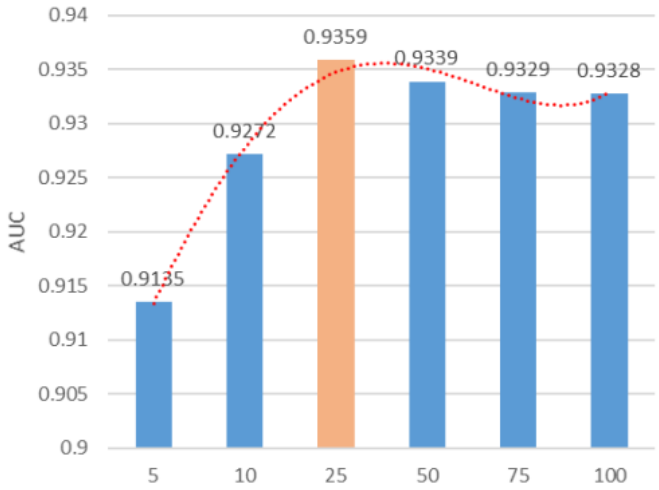

Figure. 8. Prediction performance of GCMDR by setting different numbers of latent factor model.

4.3.6. Evaluation of negative sampling's effectiveness

Previous work reports that leveraging unlabeled data in training can improve prediction performance significantly if used properly [143]. As there are only positive samples in the collected dataset, we need to find negative samples for semi-supervised training to find a prediction model. To do so, we performed sampling on the unlabeled microRNA-drug pairs to generate negative samples for training. The size of the negative datasets we used is fixed in each sampling. As part of our experiments, we evaluated the influence of the ratio, ***p***, of the size of the negative dataset to that of the positive dataset on the prediction performance. As shown in Table 4, we noted that the best prediction performance can be obtained if the size of the negative dataset was set to be 10 times that of the positive dataset. Specially, when ***p*** is set to 0, it means that no negative dataset was used and that we performed training based only on the positive samples. The huge improvement of the AUC from ***p***=0 to ***p***=10 is a strong evidence that the negative datasets are important for the training of the GCMDR. In addition, Figure 2(c) and 2(d) show that, in the training process, both of training loss and prediction error were convergent when the number of training epoch increases to about 150.

Table 10 Performance comparison w.r.t. AUC in 5-fold CV using different settings of negative sampling.

value of ***p***	10	5	1	0
Average AUC	0.9359+/- 0.0006	0.9297+/- 0.0002	0.9235+/- 0.0002	0.8083+/- 0.0111

4.4. Summary

There are more and more evidence that microRNA expression levels are related to drug

resistance. However, there is still very little understanding about how microRNA expression levels are associated with drug failure. To date, little effort has been made to attempt to predict, on a larger scale, if there exists any association between microRNA and drug resistance. In this paper, we propose a deep learning-based computational model for this purpose.

To the best of our knowledge, GCMDR is the first computational tool developed to predict the influence of microRNAs on drug resistance. Based on the results obtained, the prediction models developed using GCMDR could provide useful insights for drug design. They may also help to reevaluate the efficacy and toxicity of drugs for patients with different microRNA expression profiles. In an attempt to provide a more comprehensive dataset for future research in this field, we have publicly released the data of various attributes of microRNAs and drugs that are recorded in ncDR database. Other than making such contribution, it should be noted that GCMDR enables an end-to-end prediction for associations between microRNAs and drug resistance. This flexibility makes it possible for GCMDR to have a wide scope of application with any numerical type of intrinsic features. Finally, based on knowledge available about microRNA-induced drug resistance, it should also be noted that GCMDR is able to learn a new type of features for microRNAs and drugs based on graph embedding. The embedding features learnt by GCMDR allow a similarity measure to be derived for microRNAs with regard to their drug effects, and this is expected to help discover microRNA functional clusters using a clustering algorithm.

5. EPLMI&LMNLMI - ALGORITHMS FOR PREDICTING LNCRNA-MICRORNA INTERACTION

5.1. Background

The discovery of the essential role of non-coding RNAs (ncRNAs) in the regulation of gene expression leads many to believe that the transcriptional landscape of many organisms is far more complex than previously thought [144]. Non-coding RNAs, in the vast majority of transcripts expressed in mammals, have lengths ranging from 22 nucleotides to hundreds of kb. The *long non-coding RNA* (lncRNA) among the ncRNAs is a loosely classified group of RNA transcripts longer than 200 bases with no apparent protein-coding function and they can be found in every branch of life [13]. There has recently been increasing evidence that lncRNAs can be involved in various cellular processes, such as cell differentiation, cell growth and death, etc. They seem to be able to exert influences over chromatin modification, transcriptional complex targeting, mRNA splicing and protein translation. The past few years have witnessed a surge of interest in the development of computational tools for the identification and annotation of non-coding RNA [145-148]. However, even though more than 58,000 human lncRNA genes have been identified, apart from the few lncRNAs, like XIST and HOTAIR, that are well-studied, the role that most lncRNAs can play in different cellular processes remain largely unknown due to the complex and dynamic molecular mechanisms [149].

LncRNAs have been found to be able to regulate patterns of expressed proteins via a specific mechanism composed of different kinds of biological interactions such as the interactions between lncRNA and protein, lncRNA and mRNA, and lncRNA and ncRNA [150]. As a result, the construction of maps of putative biological interaction network mediated by lncRNAs could be necessary for the understanding of potential biological functions and mechanisms of lncRNAs. As a main kind of competing endogenous RNAs (ceRNAs), lncRNAs can function as microRNA sponges, leading to lower regulatory effect of microRNA on mRNAs, and therefore microRNAs play significant roles in the molecular mechanisms of lncRNAs [144]. Previous work on function annotation of lncRNAs are mainly based on expression correlation between lncRNAs and protein-coding genes across different tissues [151, 152]. Few functional annotations were conducted based on the ceRNA network. Given the knowledge accumulated over microRNA function for the past decade, if the interaction between lncRNA-microRNA can be better understood or even predicted, we can gain great insights into the complex functions of lncRNA.

Recently, there are more and more evidence to show that both microRNA and lncRNA are implicated in the pathological processes involved in diverse human diseases. And as a result, there has been much effort to investigate into the impacts that microRNA can have on lncRNA functions and vice versa [16, 149]. For example, lncRNA-microRNA regulatory networks in prostate cancer, gastric cancer and vascular diseases have been constructed [153-155] have been studied. Such detailed understanding of the effects of lncRNA-microRNA interactions can have in pathophysiology could pave the

way for new biomarker discovery and therapeutic approaches. Unfortunately, however, the interaction between lncRNA-microRNA as identified by biological experiments is still too limited for such understanding to make very wide impacts.

To expedite the process of identifying such microRNA-target interactions, it is common practice to perform *In silico* prediction to refine the candidate list for further validation experiments [156]. Existing computational algorithms developed for predicting such microRNA-target predictions are designed with several common rules that address the four aspects of conservation, seed match, free energy, and site accessibility [149]. However, many microRNA-target prediction tools are developed originally for mRNA targets, and as a result, predictions are made based on the nature and statistical rules of mRNA-microRNA interactions and may contradict with that of lncRNA-microRNA interactions. [157]. For example, some existing prediction methods for microRNA-target interactions perform conservation analysis focusing on the regions in the 3' UTR and the 5' UTR of mRNA based on the observation that the microRNA seed region of mRNA usually has higher conservation than the non-seed regions. However, lncRNA is reported to show significantly lower sequence conservation and evolve faster than mRNAs [149].

In addition to this, it is also noted that as the strategy of seed match is based on the statistical rules originally obtained for microRNA-mRNA interactions, they would be unsuitable for lncRNA-microRNA interaction prediction.

Besides, a few models proposed for prediction of lncRNA-RNA interaction perform

their tasks by simply computing the free energy of the potential binding sites [149]. For example, LncTar computes the free energy which is to measure the stability of complementarity between lncRNA and target RNA [157]. However, although this kind of sequence-based prediction algorithm has a wide application range, they are plagued by very high false positive rates [156]. Other than this, some inherent characteristics are found to differentiate lncRNAs from mRNAs. For example, comparing with mRNAs, lncRNAs are generally found to be shorter with fewer exons. They are also more lowly-expressed, more enriched in the nucleus, and show higher tissue-specificity and reduced stability [149]. Most existing microRNA target prediction tools fail to incorporate recent advancements in the understanding of lncRNA-microRNA interaction and may therefore not effective enough for the prediction of lnRNA/microRNA targets for a specific microRNA/lncRNA.

Recent theoretical and experimental research have shed light on the modeling of the crosstalk between different kinds of ceRNAs, including lncRNA and microRNA within the cell [158]. It appears that, apart from other well-known factors such as sub-cellular localization and microRNA response element (MRE) accessibility associated with secondary structures or RNA-binding protein, the expression levels of individual lncRNA and microRNA has come to be the key to decipher the rules of ceRNA networks [159].

Previous work on protein-protein interaction predictions [160], small RNA (sRNA) regulation [161] and microRNA-target threshold effects [162] reveal that, as the two

major components of ceRNA network, lncRNAs and microRNAs interact with each other according to a titration mechanism which orchestrates their interaction by establishing a threshold level of effect. The basic postulate of this titration mechanism is that optimal lncRNA-microRNA cross-regulation occurs at a near-equimolar equilibrium because lncRNA would be inactive in the presence of limited number of available microRNA and, conversely, be fully repressed when microRNA molecules are much more abundant [159]. In other words, RNA dosage is critical for cross-regulation and the baseline expression levels of microRNA and lncRNA can offer important insights into their direct and indirect interaction patterns according to the overall network equilibrium.

Based on such considerations, Ala et al. proposed a kinetic mathematical model to predict ceRNA interactions mediated by phosphatase and tensin homolog (PTEN). This kinetic model makes use of transcription and degradation rates for microRNA/ceRNAs association/dissociation and the degradation rates for microRNA/ceRNA complexes as the model's key parameters [159]. However, all these parameters are hard to be defined for most lncRNAs and microRNAs and as a result, the kinetic model cannot be widely used for predicting lncRNA-microRNA interactions. The result of Ala's work demonstrates that ceRNA crosstalk has a close relationship with the expression levels of relative microRNAs, and the specificity of ceRNA interactions may depends on the expression profiles of microRNA.

5.2. Data collection

For the purpose of our investigation, we obtained the February 2017 version of the lncRNASNP database which is made available for downloading at http://bioinfo.life.hust.edu.cn/lncRNASNP. The database contains information about known lncRNA-microRNA interactions confirmed by laboratory studies [163]. They were collected from 108 CLIP-Seq datasets and there are 8091 records in total. After removing the duplicated entries, we obtained 5348 of them for our experiments. These records represent lncRNA-microRNA interactions, involving 780 different types of lncRNAs and 275 different types of microRNAs, respectively.

In addition, for the purpose computing the similarities among microRNAs, we have collected three kinds of information from various databases. The first of such information is related to the interaction between microRNAs and different target genes and is obtained from miRTarBase (release 6.1, http://miRTarBase.mbc.nctu.edu.tw) [5, 164]. After matching the ids of the 275 types of microRNAs, we managed to obtain information on 272 of them.

The second type of information is obtained from the expression profile data of microRNAs. The data were downloaded from the microRNA.org database (http://www.microrna.org/microrna/home.do) where the expression profiles of 230 microRNAs were found [14]. Each record of microRNA expression profile has 172 dimensions representing the expression levels of a single type of microRNAs in 172 different human tissues and cell lines.

The third type of information is obtained from the sequence data of mature microRNAs.

The data were obtained from the miRBase database (http://www.mirbase.org/index.shtml) [26].

To compute the similarity among lncRNAs, we downloaded the putative functional annotations of lncRNAs from the NONCODE database (http://www.noncode.org/) [165]. After converting the names of lncRNA into the NONCODE IDs, we successfully obtained expression profile data for 450 of the lncRNAs and the functional annotations of 264 of the lncRNAs.

The collected expression profiles of lncRNA have 22 properties describing the expression level of each type of lncRNAs in 16 different human tissues and 8 cell lines. The functional annotations for lncRNA genes we obtained describe the 10 most probable biological functions as predicted by lnc-GFP method based on a coding–non-coding co-expression network.

Finally, for performance assessment, we have also downloaded the sequence data of lncRNAs from LNCipedia database (https://lncipedia.org/) [13].

5.3. EPLMI and LMNLMI in details

5.3.1. Motivation

There are increasing evidences that certain lncRNAs are presumably co-regulated in expression networks, suggesting that multiple lncRNAs may regulate biological processes through interacting with specific microRNA clusters in a synergistic manner

[166, 167]. It may reasonably be assumed that there is lncRNAs interacting with same microRNAs are expressed similarly across different tissues and cell lines.

Therefore, we investigated into the expression patterns of a large number of lncRNA-microRNA interactions identified by high-throughput experiments, and have discovered that the microRNAs that have been identified to interact with specific lncRNA tend to share more similar expression pattern than those are not known to be interactive. Conversely, the expression profiles of lncRNAs that have been identified to interact with the same microRNA also tend to be more similar than those of the others.

Motivated by this discovery and the limited knowledge known about MRE binding rules, we propose here a computational model to predict large-scale lncRNA-microRNA interaction network as a whole. To the best of our knowledge, this is the first of its kind.

Without using the sequence data of lncRNAs and microRNAs, we develop a two-way diffusion model called EPLMI to predict new lncRNA-microRNA interactions and compute the putative interaction strength of known lncRNA-microRNA interactions based on known lncRNA-microRNA interaction network. The basic assumption behind the development of the model is that lncRNAs of similar expression profiles tend to interact with a cluster of microRNAs having similar expression profiles, and vice versa.

Most existing microRNA target prediction methods cannot be easily adapted to take into consideration more and more newly discovered information about microRNA and

apply it effectively to predict lncRNA–microRNA interactions. To further improve the performance of EPLMI, we proposed LMNLMI model. EPLMI only uses information relating to expression profiles of microRNAs. The other kinds of information are not considered by the research team.

As its related work in other areas has demonstrated successfully the usefulness of integrating multi-domain features to accomplish the prediction task, we believe that knowledge about multimodal networks can be useful [32-33]. Some recent work has started to address multiple network integration by combining the network diffusion algorithm with dimensionality reduction scheme [34-37]. They use a subtle fusion method to fuse multiple similarity networks for interaction prediction, cancer patient clustering and Kidney Renal Cell Carcinoma identification. Despite the performance of these approaches being better in drug interaction prediction and cancer patient clustering, they are not applicable to multi-network based lncRNA–microRNA interactions prediction.

5.3.2. Construction of diverse lncRNA/microRNA similarity matrixes

Based on the assumption that lncRNA/microRNA tends to interact with a cluster of microRNAs/lncRNAs which share similar features and regulation patterns, we have investigated into three different types of lncRNA/microRNA similarity by incorporating diverse information resources. The similarity matrix we computed for the first type is based on the data of expression profiles. Specifically, we use Pearson correlation coefficient for similarity measurement. Given two expression profiles of

two RNAs (say e_a and e_b), the correlation coefficient score is computed as follow:

$$ES(a,b) = \frac{\sum_{i=1}^{N}(e_{ai} - \overline{e_a})(e_{bi} - \overline{e_b})}{\sqrt{\sum_{i=1}^{N}(e_{ai} - \overline{e_a})^2 \sum_{i=1}^{N}(e_{bi} - \overline{e_b})^2}} \tag{31}$$

where N denotes the number of properties of the expression profiles and is 172 for microRNAs and 22 for lncRNAs. A pair of RNAs with a higher correlation score is considered to be more similarly expressed in general.

The second kind of RNA similarity we used is based on putative biological functions. Based on the assumption that microRNAs targeting more of the same genes tend to be involved in similar biological functions, the data of microRNA-target gene interactions are used to measure how functionally similar each microRNA-microRNA pair is. Given two sets of target genes respectively associated with microRNA m_a and microRNA m_b (say G_a and G_b), we compute a functional similarity measure as follow:

$$FS(m_a, m_b) = \frac{card(G_a \cap G_b)}{\sqrt{card(G_a)} \cdot \sqrt{card(G_b)}} \tag{32}$$

Similarly, given two sets of putative functional annotations of two lncRNAs(say F_a and F_b), their functional similarity can be computed as follow:

$$FS(l_a, l_b) = \frac{card(F_a \cap F_b)}{\sqrt{card(F_a)} \cdot \sqrt{card(F_b)}} \tag{33}$$

To compute the sequence similarity of lncRNAs and microRNAs, we implemented the Needleman-Wunsch pairwise sequence alignment by using the package of *pairwise2* in *Biopython* [168]. Specifically, we set the identification score, gap-open penalty and

gap-open extending penalty as 2, -0.5 and -0.1, respectively [168].

5.3.3. Model structure of EPLMI

In recent years, the data of known lncRNA-microRNA interactions are being accumulated along with the development of high-throughput biotechnology, such as CLIP-seq. However, known lncRNA-microRNA interaction network is far from being completed due to the dynamic nature of the regulatory mechanism of microRNAs. Here, we propose a graph-based prediction method to infer the most potential lncRNA-microRNA interactions based on known lncRNA-microRNA interaction network, lncRNA-lncRNA similarity and microRNA-microRNA similarity. Specifically, the interaction data are represented by a bipartite graph between lncRNA and microRNA nodes, with identified interactions represented by links. The absence of a link would be considered as a potential interaction between a lncRNA and a microRNA that have not yet been experimentally confirmed. The task of lncRNA-microRNA interaction prediction can thus be mapped to predicting links in the bipartite graph of known lncRNA-microRNA interactions, labeled with prediction scores.

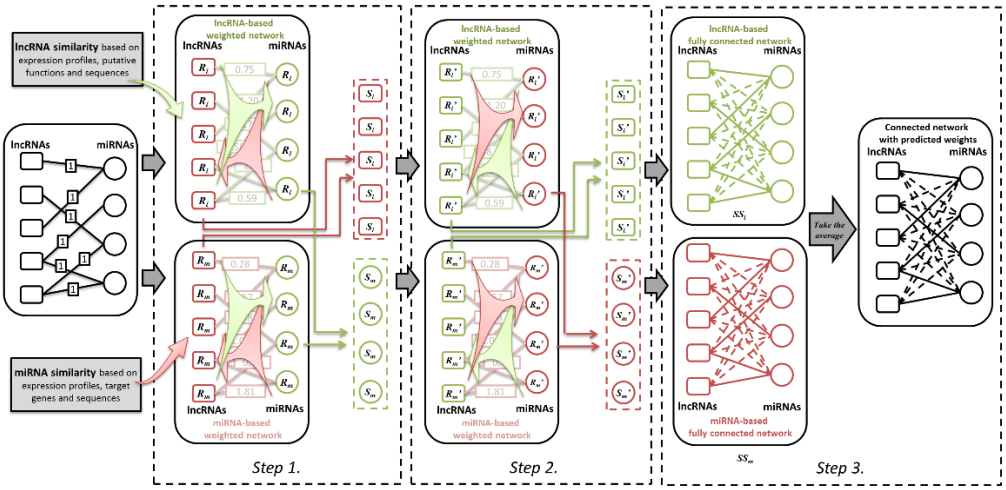

Figure 8 The flowchart of prediction process of EPLMI

In the prediction process of EPLMI, message flow forward and backward from one side of bipartite graph to another based on a two-way diffusion method (see Figure 1). Specifically, EPLMI performs its tasks in three main steps. In the first step, two kinds of weighted lncRNA-microRNA interaction networks are generated in order to introduce lncRNA/microRNA similarity into the known lncRNA-microRNA interaction network. Given the corresponding adjacency matrix $\boldsymbol{A} \in \mathbb{R}^{nl \times nm}$ of the known lncRNA-microRNA interaction network, the lncRNA similarity matrix $\boldsymbol{LS} \in \mathbb{R}^{nl \times nl}$ and the microRNA similarity matrix $\boldsymbol{MS} \in \mathbb{R}^{nm \times nm}$, two adjacency matrixes for two weighted networks is computed as follow:

$$A^l = LS \cdot A \tag{34}$$

$$A^m = A \cdot MS \tag{35}$$

where *nl* and *nm* respectively denote the numbers of lncRNAs and microRNAs in the dataset. The entity $\boldsymbol{A^l(i,j)}$ in $\boldsymbol{A^l}$ denotes the total sum of the similarity between the *i*-th

lncRNA and those lncRNAs interacting with the *j*-th microRNA. Similarly, $A^m(i,j)$ in A^m denotes the total sum of similarity between the *j*-th microRNA with those microRNAs interacting with the *i*-th lncRNA. Based on the weighted lncRNA-microRNA interaction networks, the resource vectors for both lncRNA and microRNA are further computed as follow:

$$R_{lncRNA_a} = \sum_{m=1}^{nm} \frac{A^w_{a,m} \cdot A_{*,m}}{\sum_{i=1}^{nl} A^w_{i,m}} \tag{36}$$

$$R_{miRNA_b} = \sum_{l=1}^{nl} \frac{A^w_{l,b} \cdot A_{l,*}}{\sum_{i=1}^{nm} A^w_{l,i}} \tag{37}$$

Here, Aw denotes the weighted adjacency matrixes which could be either Al or Am. We further encode the correlation be-tween one type of microRNA/lncRNA and all types of lncRNA/microRNA as a resource vector. Specifically, the resource vectors for lncRNAs, i.e. R_{lncRNA}, are actually nm-dimension row vectors and microRNA resource vectors, i.e. $R_{microRNA}$, are nl-dimension column vectors, which describe the correlation scores during forward propagation. We further set the resource vectors for step 2, i.e. S_{lncRNA} and $S_{microRNA}$, as the average of those computed based on two weighted networks. In the second step, the message flow backward to the side it starts in step 1. To obtain the correlation scores during backward propagation, the resource vectors for lncRNA and microRNA are computed based on S_{lncRNA} and $S_{microRNA}$ as follows:

$$R'_{lncRNA_a} = \sum_{m=1}^{nm} \frac{A^w_{a,m} \cdot S_{miRNA_m}}{\sum_{i=1}^{nl} A^w_{i,m}} \tag{38}$$

$$R'_{miRNA_b} = \sum_{l=1}^{nl} \frac{A_{l,b}^{w} \cdot S_{lncRNA_l}}{\sum_{i=1}^{nm} A_{l,i}^{w}} \tag{39}$$

Two types of resource vectors are further combined as the re-source vectors for the third step, i.e. $\boldsymbol{S'_{lncRNA}}$ and $\boldsymbol{S'_{microRNA}}$, by simply taking the average. In the third step, the resource vectors of lncRNA and microRNA are respectively concatenated as two nl×nm matrixes, $\boldsymbol{SS_{lncRNA}}$ and $\boldsymbol{SS_{microRNA}}$, which are correspond to two fully connect networks (see Step 3 in Figure 8):

$$SS_{lncRNA} = [S'_{lncRNA_1}{}^{T}, S'_{lncRNA_2}{}^{T}, ..., S'_{lncRNA_{nl}}{}^{T}]^{T} \tag{40}$$

$$SS_{miRNA} = [S'_{miRNA_1}, S'_{miRNA_2}, ..., S'_{miRNA_{nm}}] \tag{41}$$

As a result, the final predict network could be computed with the average of $\boldsymbol{SS_{lncRNA}}$ and $\boldsymbol{SS_{microRNA}}$ as its adjacency matrix $\boldsymbol{SS}$:

$$SS = \frac{SS_{lncRNA} + SS_{miRNA}}{2} \tag{42}$$

5.3.4. Model structure of LMNLMI

Proper integration of different types of side information is crucial for effective prediction of a computational model [35, 169-173]. Above the similarity network from multiple sources are inherently correlated, and sometimes provide complementary information to each other. As a result, network fusion has been paid much attention to, which mainly aims to generate most similar representation between entities under the existing domains. Following the criteria defined for the desirable lncRNA and

microRNA similarity networks, we now formulate a fusion problem by combining several networks. This procedure is inspired by the network learning work developed for the genome-wide data analysis [174]. In order to construct a fused network from multimodal networks, we use a normalized weight matrix $P = D^{-1}W$ as the full kernel on the vertex set V. D is a diagonal matrix that entries $D(i,i) = \sum_j W(i,j)$, and $\sum_j P(i,j) = 1$. For a better normalization, our approach looks at make this free of the self-similarities on the diagonal entries of W, and keep $\sum_j P(i,j) = 1$:

$$K(i,j) = \frac{L(i,j)}{2\sum_{k \neq i} L(i,k)} \tag{43}$$

subject to the constraints: $i \neq j$,otherwise $K(i,j) = \frac{1}{2}$.

Let N_i represent a set of v_i's neighbors including v_i in G. We then measure local affinity by using K nearest neighbors as follows:

$$Q(i,j) = \frac{L(i,j)}{\sum_{k \in N_i} L(i,k)} \tag{44}$$

subject to the constraints: $j \in N_i$, otherwise $Q(i,j) = 0$

Let $K_{t=0}^{(v)}$ represent the initial v status matrix at beginning, anthe the d $Q_{t=0}^{(v)}$ represent the kernel matrix. The fusion step is iteratively updating similarity matrix corresponding to each of the data types:

$$K^{(v)} = Q^{(v)} \times \left(\frac{\sum_{k \neq v} K^{(k)}}{m-1}\right) \times (Q^{(v)})^T, v = 1,2,3, \dots, m \tag{45}$$

This procedure updates the status matrices each time to generate v parallel

interchanging diffusion processes. After t steps, we generate the final status matrix as follows:

$$K^{(v)} = \frac{\sum_{v=1}^{m} K^{v}}{m} \tag{46}$$

Formally, let $X = \left[x_1, \ldots, x_{N_d}\right]^T$, $x_i \in R^{f_d}, i = 1, \ldots, Nd$ represent a fused network of the lncRNAs, where each row i represents the corresponding feature vector of lncRNA and Nd stand for the numbers of lncRNAs. That is to say, we can use $Y = \left[y_1, \ldots, y_{N_t}\right]^T$, $y_i \in R^{f_t}, i = 1, \ldots, Nt$ to denote the corresponding feature vector of microRNA and Nt stand for the numbers of microRNAs. In particular, $X \in R^{N_d \times f_d}$ and $Y \in R^{N_t \times f_t}$ are generated from final status matrix of the network fusion section. Let A be a lncRNA–microRNA interaction matrix, where each entry $A\ ij = 1$ if lncRNA i is known to interact with microRNA j, and $A\ ij = 0$ otherwise. To infer unknown lncRNA–microRNA interactions in A, we deploy a bilinear function to learn the projection matrix P between lncRNA space and microRNA space. Generally, the bilinear function can be defined as:

$$XPY^T \approx A \tag{47}$$

where $A \in R^{N_d \times N_t}$ denoted as the known lncRNA–microRNA interaction matrix and $P \in R^{f_d \times f_t}$, R_{fd_ft} is the projection matrix that we need to learn.

We then measure the possibility of binding each pair of lncRNA–microRNA to determine whether lncRNA i more probably interacts with microRNA j:

$$\text{score}(i,j) = \text{x}_i P \text{y}_j^T \tag{48}$$

Obviously, the higher score means a greater chance of lncRNA–microRNA will interact with each other.

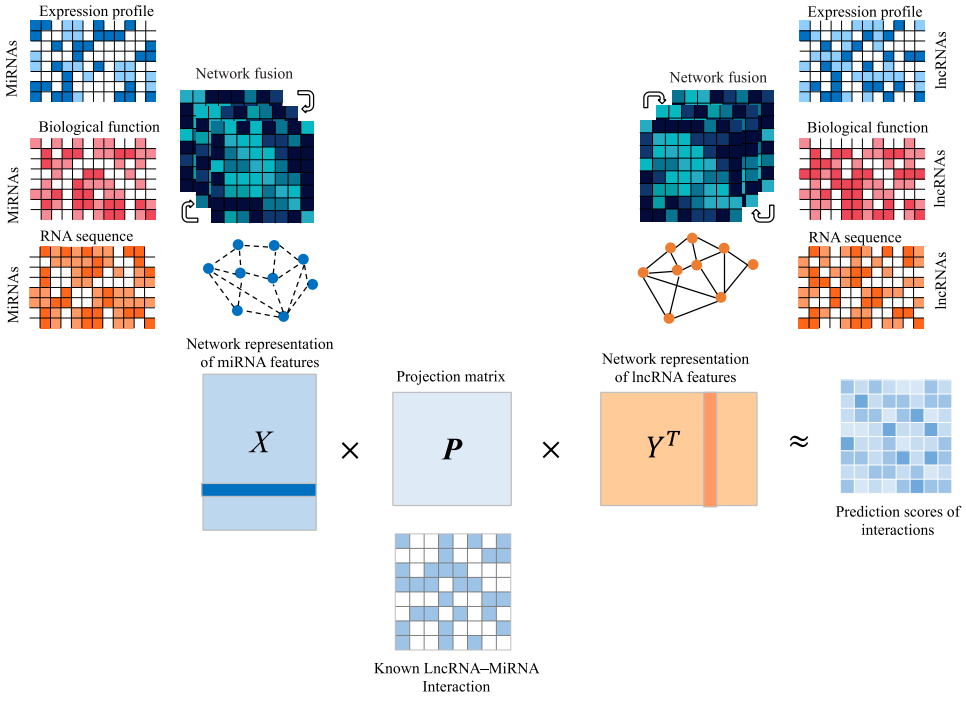

Figure 9 The flowchart of the LMNLMI pipeline. LMNLMI first integrates a variety of RNA-related information sources to construct a heterogeneous network. LMNLMI then finds the best projection from lncRNA space onto microRNA space, such that the projected feature vectors of lncRNA are geometrically close to the feature vectors of their known interacting microRNA. After that, LMNLMI infers new interactions for a lncRNA by sorting its target candidates based on their geometric proximity to the projected feature vector of this lncRNA in the projected space.

Although the projection matrix P is of dimension $f_d \times f_t$, there typically exist significant correlations between those feature vectors of lncRNAs or microRNAs that are geometrically close in space, which can thus greatly reduce the number of effective parameters required in P to model lncRNA–microRNA interactions. To take into account this issue, we impose a low-rank constraint on P to learn only a small number

of latent factors, by considering a low-rank decomposition of the form:

$$P = WH^T \tag{49}$$

where $W \in R^{f_d \times f_t}$ and $H \in R^{f_t \times f_t}$. This low-rank constraint not only alleviates the overfitting problem but also computationally benefits the optimization process [40]. The optimization problem with such a low-rank constraint on the original projection matrix P is NP-hard to solve. A standard relaxation of the low-rank constraint is to minimize the trace norm of the matrix (10), which is equivalent to minimize: $\frac{1}{2}(\|W\|_F^2 - \|H\|_F^2)$. Therefore, factoring P into W and H can be accomplished by solving the following optimization problem by alternating minimization:

$$\min_{W,H} \sum_{(i,j)} \left\| A_{ij} - x_i W H^T y_j^T \right\|_2^2 + \frac{\lambda}{2}(\|W\|_F^2 - \|H\|_F^2) \tag{50}$$

5.4. Experiment and analysis

5.4.1. Comparison of expression profiles between identified and unidentified lncRNA-microRNA interactions

For the purpose of assessing the effectiveness of EPLMI, we have investigated into the differences in the correlation of the expression profiles between identified and unidentified lncRNA-microRNA interactions. Based on known lncRNA-microRNA interaction network, we compared the differences in the expression profiles of two groups of microRNA/lncRNA pairs: (i) connected and (ii) unconnected microRNAs/lncRNAs for each single lncRNA/microRNA.

For each microRNA node that has more than two links, we divide the lncRNAs into two groups which we refer to as the identified microRNA group and the unidentified microRNA group, according to whether it is identified to interact with the microRNA or not. For each of the two groups, we computed the average Pearson correlation coefficients (PCC) of the expression profiles in the group.

For the purpose of comparison, we used the average PCC of the unidentified group as the baseline score for each lncRNA. We noted as a result that approximately 83.50% of the lncRNAs (435/521) tend to cooperate with a cluster of microRNAs sharing more similar expression profiles than the baseline (see Figure 10). For the 521 types of lncRNAs, the average PCC value of their identified microRNA groups reaches 0.4947, which is significantly higher than the average baseline value of 0.4551. In addition, if we are to highlight the samples having significantly higher or lower PCC than the baseline by using a difference threshold of 0.5 times standard deviation of PCC of identified microRNA groups (i.e., 0.058), we can see that 80.16% (202/252) of marked samples higher than the baseline (see Figure 10).

Considering that a number of lncRNAs expression profiles are unavailable and that the microRNAs in our dataset are found to have interaction with an average of approximately 19 types of lncRNA, we therefore focus only on those 206 well-studied microRNAs that have more than 5 links in order to obtain more reliable conclusions.

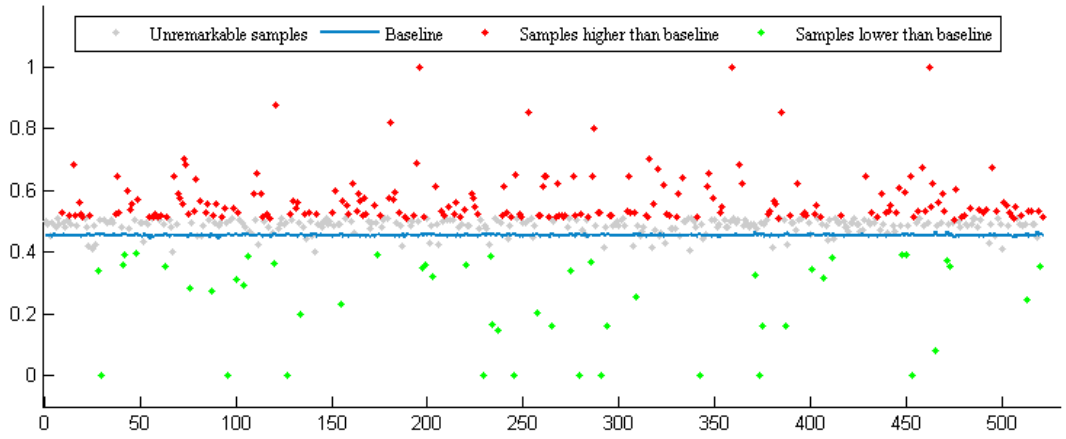

Figure 10 Correlation of microRNA clusters interacting with single lncRNAs

By similar analysis with both the identified lncRNA group and the unidentified lncRNA group for each single microRNA, we found that approximately 59.22% (122/206) of the microRNAs tend to interact with a cluster of lncRNAs that have more strongly correlated expression profiles than the baseline (see Figure 11).

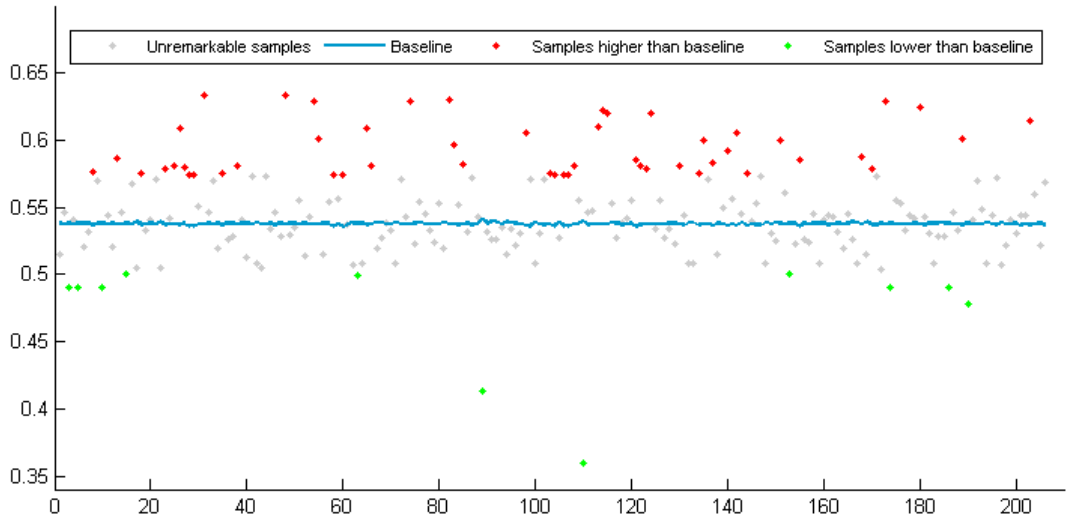

Figure 11 Correlation of lncRNA clusters interacting with single microRNAs

The average PCC of the identified lncRNA groups of 206 samples is 0.5476, which is higher than that of the baseline of 0.5378. The outstanding samples which have a different standard deviation of PCC of the identified lncRNA groups (i.e., 0.0368) from

the baseline can be highlighted and shown in Figure 10. Approximately 82.26% (51/62) highlighted samples were found to be consistent with our assumption that the target lncRNAs of a specific microRNA tend to have similar expression level patterns among different tissues and cell lines.

These results confirm the influence of expression profiles of both lncRNAs and microRNAs on their pairwise interactions. Specifically, lncRNA molecules tend to interact with a cluster of microRNAs that have similar expression profiles. In addition, we also found that most of the microRNAs are targeted by lncRNA clusters which have similar expression profiles when setting a difference threshold. However, it is noteworthy that, without setting a different threshold, only less than 60% microRNA samples are consistent with the conclusion we made. The reason of this relatively small percentage may lie in the recent finding that lncRNA displays high natural expression variation among different individuals and therefore the expression profile data we obtained may be unrepresentative [175]. Besides, due the general lncRNA feature of high tissue-specific expression, 22 dimensions of the explored lncRNA expression profile data may not be enough for comprehensively describing the expression patterns of a single lncRNA.

To further evaluate the correlation patterns of lncRNA and microRNA with respect to other kinds of lncRNA/microRNA similarity patterns, an analogous analysis was also carried out with the functional and sequential similarities. We regard those samples obtaining higher correlation scores than the baseline as positive samples that are

consistent with the basic assumption of EPLMI. As a result, 33.33% microRNA samples and 56.13% lncRNA samples are positive in the functional similarity-based experiment while 51.78% microRNA samples and 89.36% lncRNA samples are positive in the sequence similarity-based experiment.

5.4.2. Performance evaluation for EPLMI

5.4.2.1. Cross validation experiments

To evaluate the accuracy of the prediction models built by EPLMI, we used a real dataset involving confirmed lncRNA-microRNA interactions and tested accuracy using the two methods of *LOOCV* and *5-fold cross validation*.

Specifically, according to LOOCV, each known lncRNA-microRNA interaction was left out, in turn, for testing and the rest of the known lncRNA-microRNA interactions were used as training samples to construct a prediction model. To avoid the denominators in formula (36), (37), (38), (39) becoming zeros, we replace all zeros in $\boldsymbol{A}^{w}$ with a tiny value of 10^{-11}. For the purpose of deciding if the testing sample is positive, we try to compare it with the other lncRNA-microRNA pairs in the dataset whose interactions are un-confirmed. To do so, we sorted these pair samples and determined the rank of the testing sample among all the 209152 unidentified samples. If it obtains a higher rank than a given threshold, the testing sample would be considered positive.

For each different threshold set in the experiments, we obtained corresponding true positive rates (TPR, sensitivity) and false positive rates (FPR, 1-specificity) where the

sensitivity and specificity denote the percentage of testing samples with respectively higher and lower ranks than the given thresholds. In addition, we also obtained the ROCs (Receiver Operating Curves) by plotting TPR versus FPR at different thresholds and computed the values of the AUCs. The AUC values lie between 0.5 to 1 where 0.5 denotes a purely random prediction and 1 denotes a perfect performance. The best prediction model built by EPLMI achieved a reliable prediction performance with AUC of 0.8522.

Using the *5-fold cross validation*, all known lncRNA-microRNA interaction data were randomly divided into 5 subsets of roughly the same size and in each of a series of experiments, 4 would be used as training samples and the remaining data subset was used as testing samples. As was the case with LOOCV, we obtained the ROC curve for each round of 5-fold cross validation and computed the average value of the AUC. To avoid any bias caused by random partitioning of data subsets, we repeated the random sampling of data 50 times. As a result, we found that the best prediction model obtained by EPLMI achieved an average AUC of 0.8447±0.0017. We anticipate that those candidates with higher ranks would be confirmed by the experimental observation in the future. And we assume that those lncRNA-microRNA pairs that share a tight relationship in their regulation network tend to be more stable and are, therefore, more competitive in nature.

5.4.2.2. Comparison among different kinds of RNA-similarity

Apart from the expression profiles, there are other kinds of information, such as target

genes of microRNAs, putative biological functions and nucleotide sequence data, which help to describe the features of lncRNA and microRNA. In this section, we further explore two types of such information which are related to RNA-similarity: (i) RNA functional similarity and (ii) RNA sequence similarity. With EPLMI, they can be used to predict lncRNA-microRNA interactions.

To evaluate their usefulness for such purpose, LOOCV and 5-fold cross validation were implemented in this comparison experiments and the results are analyzed and discussed here (see Figure 12 and Table 11).

With regard to (i), recent efforts have been made to predict the biofunctional roles of ncRNAs but the results remain to be categorically proven. To avoid any bias in the prediction of microRNA functions, we used the data of microRNA-target gene associations to measure how functionally similar two microRNAs are. As with the functional similarity of lncRNAs, we simply followed the function annotations of lncRNA based on predictions made by previous work [176].

Table 11 Performance comparison among three kinds of RNA similarity by using EPLMI in the framework of 5-fold cross validation.

Expression profile-based similarity	Biological function-based similarity	RNA sequence-based similarity
0.8447±0.0017	0.7608±0.0011	0.7890±0.0014

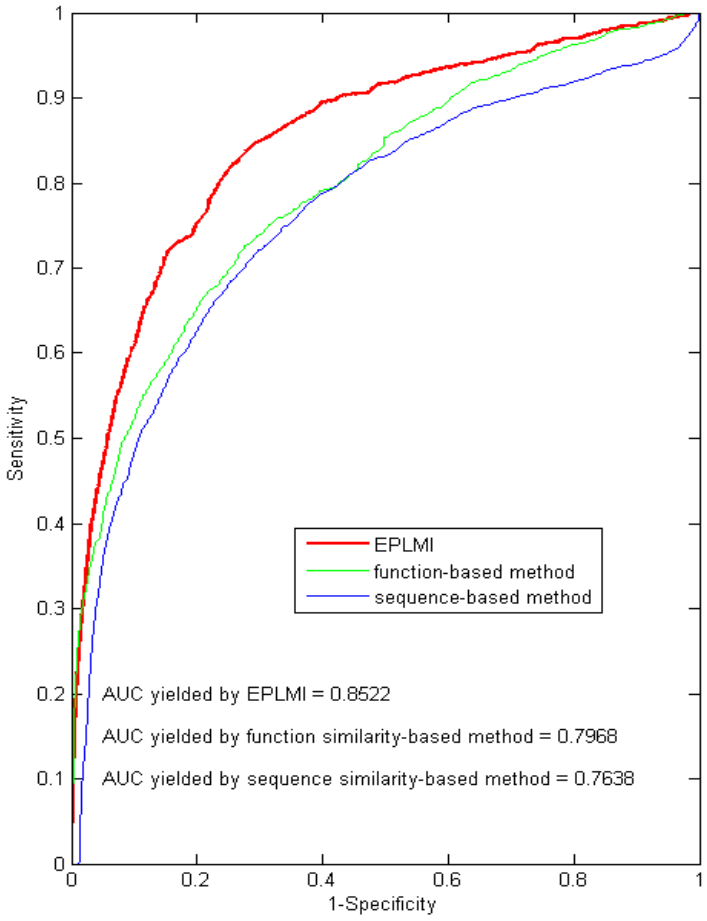

Figure 12 Performance comparison among three kinds of RNA similarity, i.e. expression profile-based, biological function-based and sequence-based similarities, by using the method of EPLMI

From the results, the prediction models built by EPLMI using RNA functional similarity and RNA sequence similarity yielded, respectively, AUCs of 0.7968 and 0.7638 in LOOCV experiments. For the *5-fold cross validation* experiments, we obtained average AUCs of 0.7608±0.0011 and 0.7890±0.0014 by using RNA functional similarity and sequence similarity, respectively.

The results of both leave-one-out and 5-fold cross validation demonstrate that the uses of functional similarity and sequence similarity of RNAs are less effective than the use of similarity based on expression profiles. The reason may lie in the fact that the

biological roles played by lncRNAs can be so diverse and many lncRNAs may not have appreciable functions which can be described by the known annotations. Hence, the putative lncRNA functional similarity based on coding–non-coding co-expression network may not be accurate and comprehensive enough for this measurement. Furthermore, the size of lncRNA sequences can be very different and the length of the lncRNAs used in this work range from 73 to 59462. For this reason, simply implementing pairwise sequence global alignment by using a dynamic programming algorithm may not be effective in the measuring of how biologically similar two lncRNAs are or how similar the regulation patterns of two lncRNAs would be. This is because microRNAs are usually sequestered by small binding sites in lncRNA.

Besides, there are increasing evidence that the expression of lncRNAs is tightly regulated and their expression profiles are important markers for the developmental stage and the disease state. Considering this noteworthy feature of lncRNA, the information of lncRNA expression profiles is considered useful for effectively depicting the correlation of lncRNAs in their microRNA-mediated regulation patterns.

5.4.2.3. Comparison with different prediction methods

To further evaluate the performance of EPLMI, we compared it with some classical prediction methods by using the same expression profile-based similarity. As the models built by EPLMI uses a network-based method through two-way diffusion, we here explore another kind of network-based method, the Katz measure, which is initially proposed for link prediction problem in social network and extensively used in

a diversity of bioinformatics problems.

Further, as the prediction task in this work can be solved as a matrix-completion problem, two main kinds of recommendation algorithms were further investigated. Specifically, two kinds of memory-based collaborative filtering (i.e., lncRNA-based CF and microRNA-based CF) and two kinds of model-based methods (i.e., singular value decomposition and latent factor model) were implemented for the prediction of lncRNA-microRNA interactions.

Table 12 Performance comparison among different methods by using RNA expression profile-based similarity in the framework of 5-fold cross validation.

Method	5-fold cross validation
lncRNA-based CF	0.6359±0.0024
microRNA-based CF	0.8235±0.0015
SVD-based method	0.4967±0.0340
Katz-based method	0.7439±0.0017
Basic latent factor model	0.8253±0.0024
EPLMI	0.8447±0.0017

From the experimental results, it is noted that EPLMI model yielded the best performance among six different algorithms we adopted for comparison using the LOOCV and 5-fold cross validation methods. Specifically, lncRNA-based CF, microRNA-based CF, singular value decomposition (SVD), latent factor model (LFM) and Katz method respectively yielded AUCs of 0.6452, 0.8307, 0.5009, 0.8271 and 0.8073 in LOOCV, and average AUCs of 0.6359±0.0024, 0.8235±0.0015, 0.4967±0.0340, 0.8253±0.0024 and 0.7439±0.0017 in 5-fold cross validation (see Figure 13 and Table 12). Compared with the other algorithms, the outstanding performance of EPLMI demonstrates that it has reliable prediction performance for

large-scale lncRNA-microRNA interactions by well incorporating the information resources of expression profiles.

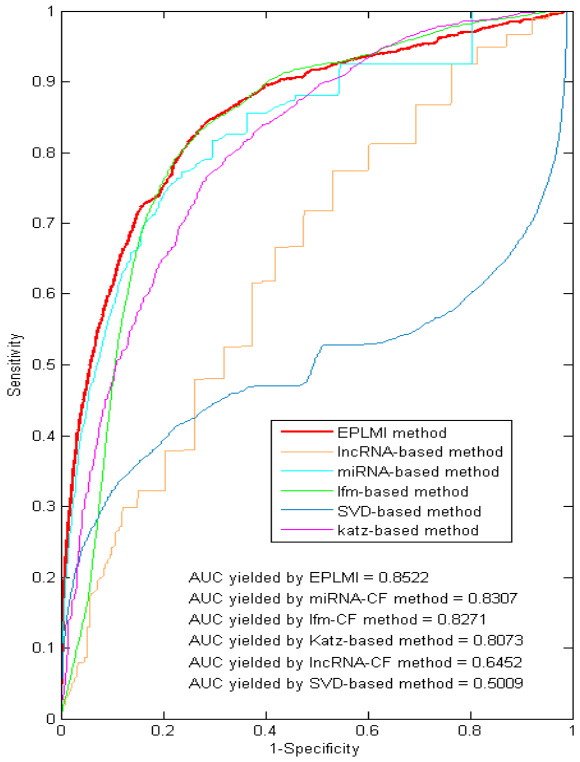

Figure 13 Performance comparison of EPLMI with 5 different kinds of classical methods by using the same RNA expression profile-based similarity.

5.4.3. Performance evaluation for LMNLMI

5.4.3.1. Cross validation experiments

Regarding prediction performance evaluation involving imbalanced data, it is suggested that ranking measures such as AUROC (area under the ROC curve) can be used so that the prediction performance can be evaluated unbiased[177]. We infer interactions and compare against the held-out interactions, measuring performance using the AUC for our evaluations. ROC curves are created by plotting the true positive rate versus the false positive rate at various thresholds. The results are shown as a ROC

curve where TPR is plotted against FPR, calculated as follows: $TPR = TP/TP + FN$, $FPR = FP/FP + TN$, where TP (true-positives) is the number of correctly predicted drug-target interactions while the FP (false-positives) is the number of not correctly predicted lncRNA–microRNA interactions. TN (true negative) is the number of lncRNA–microRNA interactions predicted not to be in a class that are not observed in that class; and FN (false negative) is the number of lncRNA–microRNA interactions predicted not to be in a class that are observed in that class. To further evaluate the performance of the proposed method, we also use the AUPRC (area under the PR curve) to evaluate its performance under different similarity network combinations. Precision is the number of correctly predicted interactions divided by the number of all returned results. Recall is the number of correctly predicted interactions divided by the number of results that should have been returned. Precision and recall are then defined as: $Precision = TP/TP + FP$, $Recall = TP/TP + FN$. Due to our interactions dataset is the high-class imbalance, we used 5-fold cross validation, where each fold leaves out 20% of the positive and negative samples for testing. Since the discovered positive samples are too small, this may lead to imbalanced bias if the dataset is randomly divided. We randomly sampled known interactions and negative pairs and divide both into each fold equally. In 5-fold cross-validation, all the known lncRNA–microRNA interactions were randomly divided into five equal parts without any overlap between any two of them. Each part was selected in turn as the test samples and the remaining four as training samples. Similarly, all lncRNA–microRNA pairs without known interactions were considered as the candidate samples. Then, the scores of test

samples and the candidate samples were computed. We compared the score of each test sample with the scores of candidate samples in turn. The prediction was considered to be successful only when the rank of test sample exceeded the given threshold value.

5.4.3.2. Performance Comparisons

To evaluate the performance of LMNLMI, we also made use of some classical approaches which include the Katz measure [54], memory-based collaborative filtering (CF) [16] and latent factor model (LFM) to predict lncRNA–microRNA interactions based on the use of the similarity networks constructed. We compared the performance of CF with LFM because our prediction step adopted the matrix completion method and this is why we compared performance of LMNLMI with the two recommendation algorithms.

To the best of our knowledge, LMNLMI is the only network fusion method developed to predict lncRNA–microRNA interactions. To further evaluate the performance of LMNLMI, we also compared it with the EPLMI which is so far the only approach proposed to predict lncRNA–microRNA interactions. Katz measure as a special algorithm to solve the problem of network link prediction, here is also used to carry out the contrast test.

Table 13 reports the scores of different algorithms on the same dataset. The resulting Auroc for LMNLMI, EPLMI, lncRNA-based CF, microRNA-based CF, KATZ and LFM on original dataset are 0.8929, 0.8402, 0.6382, 0.8215, 0.7435 and 0.8257,

respectively. ILLMI is has a 6% improvement on ROC score compared with the second best.

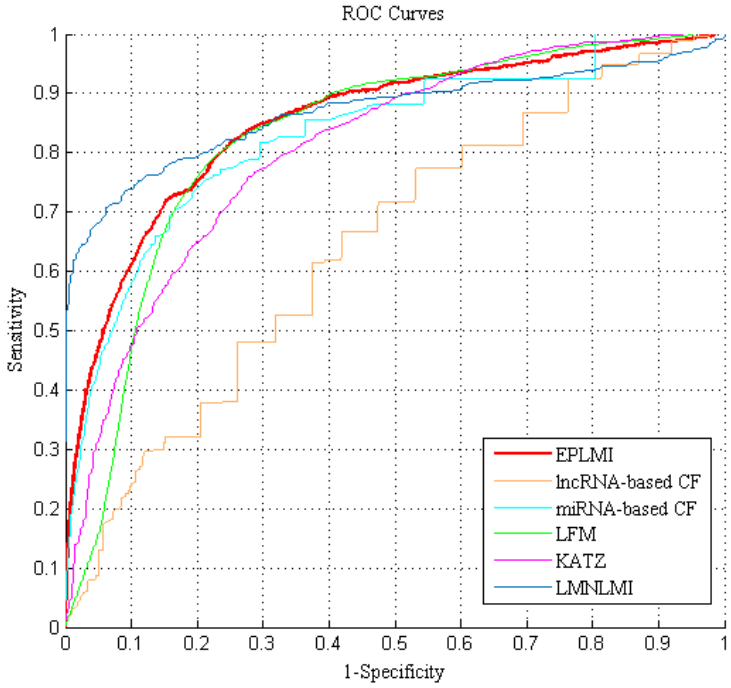

Figure 14 Comparison of the ROC curves of LMNLMI with five different kinds of methods on collected lncRNA–microRNA interactions dataset. Performance comparison of EPLMI with five different kinds of classical methods

We further compared the performance of each method by the ROC curve. Figure 14 shows the ROC curves of the six algorithms compared against the standard. As expected, among all approaches, LMNLMI achieves the highest score. This result show that LMNLMI extracted more meaningful representations to lncRNA and microRNA from fused network and this approach is shown here to have the potential to improve prediction performance.

Table 13 Performance comparison among seven kinds of RNA similarity by using LMNLMI in the framework of 5-fold cross validation.

Networks	AUROC	AUPR

Expression+Biological function+ RNA sequence	**0.8926**	**0.9223**
Expression+Biological function	0.8606	0.8980
Expression+RNA sequence	0.8670	0.9047
Biological function+RNA sequence	0.8669	0.9042
Expression-based	0.5905	0.6266
Biological function-based	0.6071	0.5930
RNA sequence-based	0.7149	0.8162

Table 14 Performance comparison among different methods by using similarity network in the framework of 5-fold cross validation.

Method	AUROC
LMNLMI	**0.8926**
EPLMI	0.8402
lncRNA-based CF	0.6382
microRNA-based CF	0.8215
KATZ	0.7435
LFM	0.8257

5.4.3.3. Cross-Test for Dataset Setting

To show the difference between the original similarity networks and fused network, the matrix representations of the similarity network that is constructed based on RNA functional information, the similarity network that is constructed based on RNA sequence information and the fused network are shown in Figure 15. In this subsection, we also conduct a series of network cross-test on three kinds of similarity networks to evaluate the impact of network fusion in LMNLMI. We tested the combination of similarity network that include the following lists: (Expression+Biological function+RNA sequence, Expression+Biological function, Expression+RNA sequence, Biological function + RNA sequence}

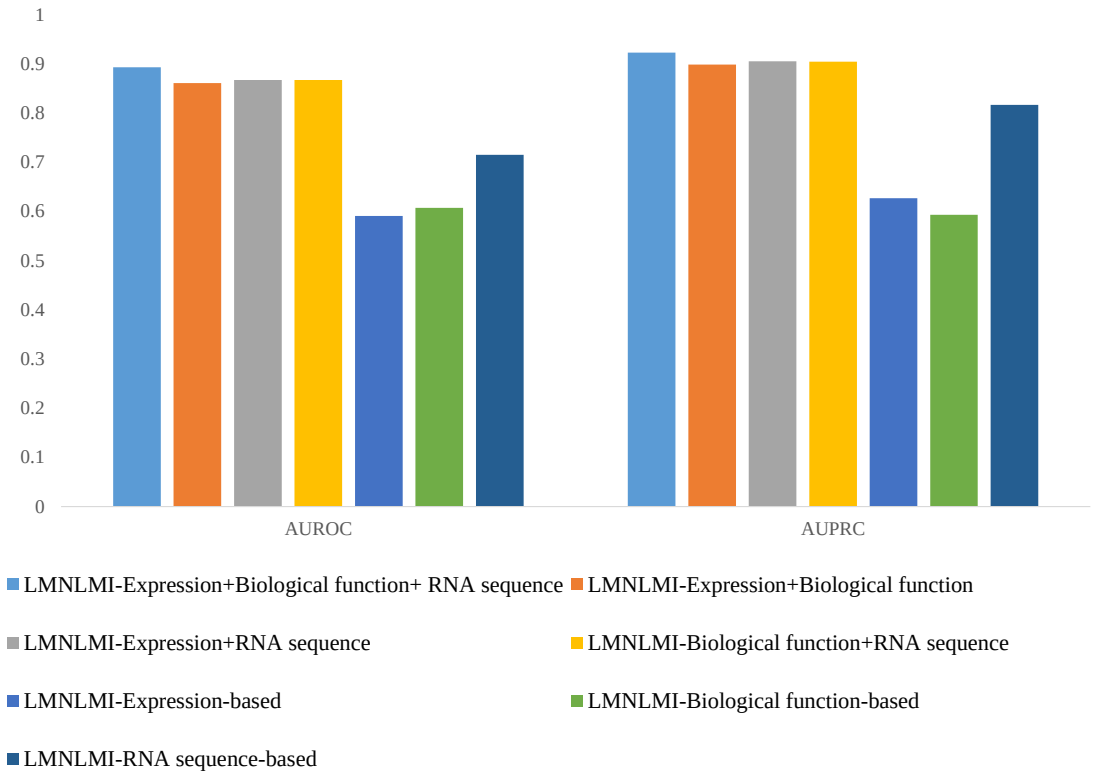

Figure 15 Performance comparison was assessed by both the area under AUROC and AUPRC among seven kinds of similarity combination, i.e. Expression+Biological function+ RNA sequence, Expression+Biological function, Expression+RNA sequence, Biological function+RNA sequence, Expression-based, Biological function-based, RNA sequence-based, by using the method of LMNLMI

Each single network is also tested with the same evaluation criterion. Performance of LMNLMI on each network combination was assessed by both the AUROC and the area under precision-recall curve (AUPRC). All results are summarized in Table 13, from which the prediction models built by LMNLMI using Expression+Bological function+RNA sequence network, Expression+Biological function network, Expression+RNA sequence network, Biological function + RNA sequence network yielded, respectively, AUCs of 0.8926, 0.8606, 0.8670 and 0.8669 in 5-fold experiments. We obtained average AUCs of 0.5905, 0.6071 and 0.7149 by using RNA Expression-based similarity Biological function-based similarity and RNA sequence-

based similarity, respectively. The results that we have obtained, as shown in the above tables, show that LMNLMI can be a promising approach for predicting lncRNA and microRNA interactions.

After evaluating the effectiveness and robustness of LMNLMI, we calculate the predicted score of interaction for lncRNA with the lowest number of known interactions in the dataset. In the real world, this kind of interaction is the most difficult to find, due to the lack of known samples. To further evaluate the ability of LMNLMI to predict the rare lncRNA-microRNA interactions, we analyzed the top 10 with the highest ranking of such lncRNA in the test set. As a result, shown in Table 15, four interacted lncRNA-microRNA pairs are finally confirmed. This result suggests that even for some ncRNAs that have not been studied thoroughly, our approach can be used to predict potential interactions.

Table 15 The top 10 predicted interactions for rare lncRNA

lncRNA	microRNA	Evidence	Score
lnc-COX10-3:1	hsa-miR-4458	lncRNASNP	0.404
lnc-COX10-3:1	hsa-miR-4500	lncRNASNP	0.403
lnc-MEP1B-1:1	hsa-miR-208a-3p	unconfirmed	0.120
lnc-IDS-1:6	hsa-miR-195-5p	unconfirmed	0.080
lnc-AEBP1-1:1	hsa-miR-520b	unconfirmed	0.075
lnc-FAM186B-1:1	hsa-miR-520b	unconfirmed	0.074
lnc-FRG2C-1:2	hsa-miR-103a-3p	lncRNASNP	0.074
lnc-C6orf164-1:1	hsa-miR-103a-3p	lncRNASNP	0.074
lnc-KLRF1-1:1	hsa-miR-4306	unconfirmed	0.072
lnc-IGSF21-2:1	hsa-miR-3619-5p	unconfirmed	0.071

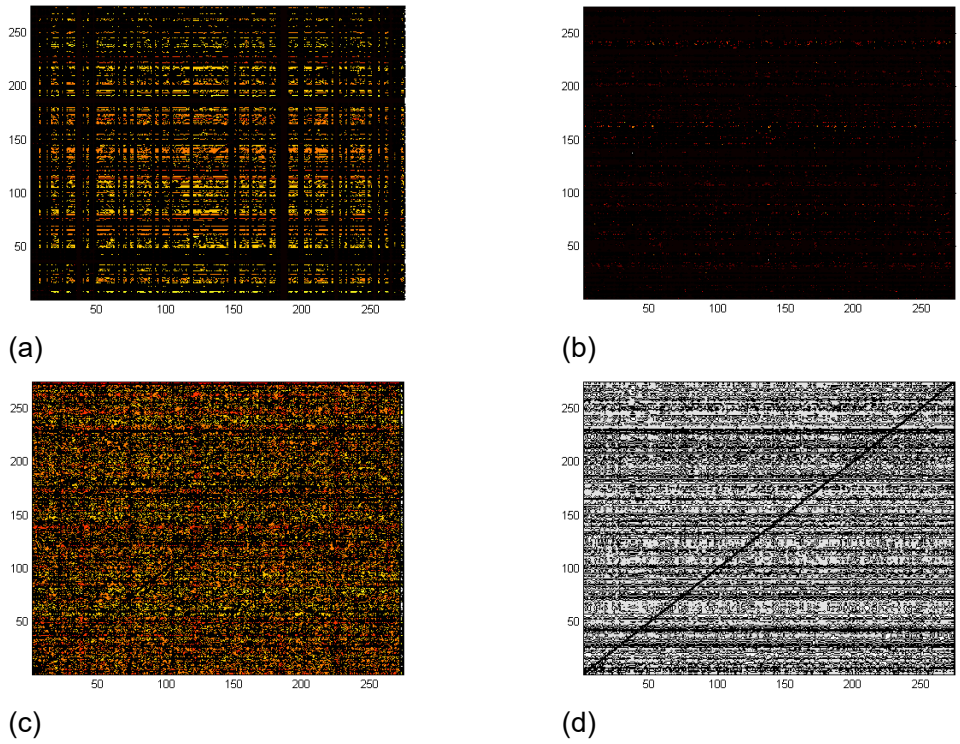

Figure 16 (a) LncRNAs Expression profile-based similarity network; (b) Biological function-based similarity network; (c) RNA sequence-based similarity network; (d) Fused similarity network based on above networks.

5.5. Summary

Even though lncRNA-microRNA interactions is becoming known to be very important for dissecting various bio-mechanisms, current knowledge and data on lncRNA-microRNA interaction that have been identified is still limited. Apart from a few sequence-based microRNA target prediction tools that mainly follow the prediction of target genes/mRNA, little effort has been made to predict lncRNA-microRNA interactions on a large scale. Based on accumulating experimental observations, the close relationship between the interaction patterns of ceRNAs and their relative expression levels has been highlighted. In this work, motivated by recent advances in the synergistic actions of lncRNAs, we analyzed statistically the patterns of large scale

lncRNA-microRNA interaction network in the perspective of expression profiles. Consequently, we discovered that lncRNAs/microRNAs interacting with the same single microRNAs/lncRNA tend to have similar expression profiles. Based on this finding, we propose the first computational technique, EPLMI, to build models for predicting large-scale lncRNA-microRNA interaction network based on a novel graph-based diffusion algorithm. The basic assumption made by EPLMI is that lncRNAs with similar expression profiles tend to collaboratively interact with microRNAs with similar expression profiles, and vice versa. By using the latest dataset of lncRNA-microRNA interactions, the experimental results obtained with EPLMI, along with a series of comparison results, demonstrate that it can be a very reliable method.

We believe that EPLMI can yield important insights into future research on ceRNA regulation networks. Unlike traditional prediction tools for microRNA-mRNA interactions, EPLMI do not focus on binding sites of microRNA in target RNAs considering that the number of binding rules of MREs are still very limited due to naturally imperfect pairing and that purely computing free-energy could yield a high rate of false positives. Instead, EPLMI predicts lncRNA-microRNA interactions by making use of the collaborative effects of both lncRNAs and microRNAs and the similarities of lncRNAs and microRNAs. By using the expression profiles of lncRNAs and microRNAs, EPLMI can yield the interaction possibility for each lncRNA-microRNA pair in one-shot and it can, therefore, have a wide-range of applications.

In addition to the above, EPLMI can offer preliminary knowledge for two other

prediction problems that we propose for future work. The first one is to predict the indirect lncRNA-lncRNA interactions. It is reported that indirect interactions occur frequently in ceRNA network where two ceRNAs can crosstalk via a third transcript. As EPLMI focuses on the common pattern of lncRNAs interacting with the same single microRNAs, the lncRNAs predicted to interact the same microRNAs with high scores may tend to have an indirect interaction.

The second one is to measure how competitive the lncRNAs are to sequester a specific kind of microRNA. Target lncRNAs may coexist as competing ceRNAs and the effectiveness and number of their MREs are not always equal, leading to different competitive status. By implementing EPLMI, those links of known lncRNA-microRNA interactions that have bigger weights could be considered as more common and biologically important than the others and therefore the lncRNAs in these interactions may have higher priority to interact with the microRNAs in order to remain biologically stable. In other words, for the known lncRNA-microRNA interactions, the lncRNAs obtaining higher scores predicted by EPLMI may be more competitive in their interaction with microRNAs.

Despite the effectiveness of EPLMI as discussed above, it should be noted that EPLMI has some limitations. As EPLMI makes prediction mainly based on datasets with known lncRNA-microRNA interaction, it may suffer from possible prediction-bias caused by imbalanced learning samples. lncRNA/microRNA that are well-studied tend to obtain a higher prediction scores since they have more links in known lncRNA-

microRNA interaction network. In addition, it should also be noted that EPLMI is not applicable to new types of lncRNA/microRNA that are without expression profiles as they do not have any links in known lncRNA-microRNA interaction network.

LMNLMI that can be used for network fusion for prediction of lncRNA–microRNA interactions is proposed. LMNLMI addresses several key challenges in biological network incompletion and multiple biological network fusion. First, it can be used to fuse diverse heterogeneous information embedded in network data. Second, it reduces the incompleteness resulting from the vertex features in the heterogeneous network data not being fully discovered. LMNLMI introduces various similarity measurements, which are used to characterize valuable information of each individual network. It then applies an SNF algorithm to the multiple networks. In addition, LMNLMI use inductive matrix completion for predicting lncRNA–microRNA interactions based on the fused network. We have demonstrated that LMNLMI has excellent ability in network integration for accurate lncRNA–microRNA interactions, inferring and achieving substantial improvement over the advanced approach. Experimental results on the real-world dataset demonstrate that LMNLMI is able to achieves a good performance.

LMNLMI is a hard-fusion method. It does not consider the weights of different networks within the fusion procedure. For future work, we plan to explore how the influence of each network can be learned directly from the data. We intend to investigate into the impacts that the degree of membership of each similarity matrix may have on the performance of predicting. In addition. Although the focus of this work is about

lncRNA and microRNA network learning, the proposed LMNLMI is flexibly connected to other biological networks, e.g. microRNA-disease associations. We will make use of LMNLMI to perform different learning tasks in different applications.

6. CONCLUSION

6.1. Summary

In this thesis, the topic we concern about is the link prediction problem in graph analytics. We propose to use machine learning techniques to solve the challenges existing in the state-of-the-art link prediction algorithms on heterogeneous attributed networks. Specifically, we concentrate on the application on one import research domain of bioinformatics, that is about link prediction for microRNA-mediated biomolecular networks. Like many other relational data in real world, the data in this domain can be formed as graphs which are often large-scale, complex and incomplete. To accelerate the research of microRNA, we develop four prediction tools for three different important research issues, each of which faces different challenge. To predict association between microRNA and human complex diseases, we aim to address the data incompleteness problem existing in the microRNA-lncRNA co-regulation network. The proposed computational model, MVMTMDA to solve this problem is based on multi-view multi-task learning and adopt a deep learning model. The second issue is to predict the kind of drug resistance associated with aberrant expression of microRNAs. To tackle the problem derived by the high dimensional attributes of nodes (drug structural fingerprint), we introduce the spectral graph convolution operator into the model of deep autoencoder model. The third prediction task in this thesis is to predict lncRNA-microRNA interactions. Different from the conventional algorithms for microRNA target prediction which are based on sequence matching, we reformulate

this problem as link prediction on graphs. We propose two methods to solve this problem. One is EPLMI which uses the information of expression profile and adopts a two-way diffusion method for prediction. The other one is LMNLMI which adopts the similarity network fusion technique and thus can effectively consider multiple type of information.

To show the effectiveness and efficiency of these proposed algorithms, we have collected real datasets from public databases and apply proposed methods to perform prediction for performance evaluation. In addition, we also compare some state-of-the-art approaches with he proposed models for further evolution. We test the robustness of proposed models in different performance with different sizes of training, parameter setting as well as different model structures. Case study was also conducted to illustrate the efficiency of the proposed model in real application scenario. The experimental results show that our proposed can be served as a useful tool for different domains of microRNA research.

6.2. Future work

In future, we attempt to improve the prediction performance of our models from two main aspects. The first direction is to consider more complex information data. When collecting the datasets, we actually found some data that are biologically relevant but unable to be utilized as input data for model prediction due to their complex data types. In the model of MVMTMDA, we have adopted multi-view multi-task learning to handle graph data. However, there are other kind of nonnumerical data in biological

research including, for example, DNA sequence that are characters of unequal length. In the field of bioinformatics, the available data are very valuable but often limited in amount. Introducing more relevant data into prediction models have much potential to improve the prediction performance. To achieve this, we will try to adopt the advanced deep learning techniques, e.g., LSTM, to handle the nonnumerical data, and to propose new feature extraction method to use prior knowledge to transform them into numerical one. The second direction of our future work is to lower the effect from data noise. The dataset we collected is from the discovery of numerical biological experiments and thus inevitably has high noise. To solve this problem, we will introduce effective dimension reduction techniques into our model. In addition, we will also investigate how to collectively use data from different domain to train a model such that the noise in one kind of data can be counteracted when fused with other kinds.

Reference

1. Langville AN, Meyer CDJIM: **Deeper inside pagerank**. 2004, **1**(3):335-380.
2. Calin GA, Croce CM: **MicroRNA signatures in human cancers**. *Nature reviews cancer* 2006, **6**(11):857.
3. Li J-H, Liu S, Zhou H, Qu L-H, Yang J-HJNar: **starBase v2. 0: decoding miRNA-ceRNA, miRNA-ncRNA and protein–RNA interaction networks from large-scale CLIP-Seq data**. 2013, **42**(D1):D92-D97.
4. Schmidt MF: **Drug target miRNAs: chances and challenges**. *Trends in biotechnology* 2014, **32**(11):578-585.
5. Chou C-H, Chang N-W, Shrestha S, Hsu S-D, Lin Y-L, Lee W-H, Yang C-D, Hong H-C, Wei T-Y, Tu S-J: **miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database**. *Nucleic acids research* 2015, **44**(D1):D239-D247.
6. Zou Q, Li J, Song L, Zeng X, Wang G: **Similarity computation strategies in the microRNA-disease network: a survey**. *Brief Funct Genomics* 2016, **15**(1):55-64.
7. Dai E, Yang F, Wang J, Zhou X, Song Q, An W, Wang L, Jiang W: **ncDR: a comprehensive resource of non-coding RNAs involved in drug resistance**. *Bioinformatics* 2017, **33**(24):4010-4011.
8. Zhang Y, Wang J: **MicroRNAs are important regulators of drug resistance in colorectal cancer**. *Biol Chem* 2017, **398**(8):929-938.
9. Bruna J, Zaremba W, Szlam A, LeCun Y: **Spectral networks and locally connected networks on graphs**. *arXiv preprint arXiv:13126203* 2013.
10. Wang JZ, Du Z, Payattakool R, Yu PS, Chen C-F: **A new method to measure the semantic similarity of GO terms**. *Bioinformatics* 2007, **23**(10):1274-1281.
11. Gillis A, Stoop H, Hersmus R, Oosterhuis J, Sun Y, Chen C, Guenther S, Sherlock J, Veltman I, Baeten J: **High-throughput microRNAome analysis in human germ cell tumours**. *The Journal of pathology* 2007, **213**(3):319-328.
12. Huang Y-A, Chan KC, You Z-HJB: **Constructing prediction models from expression profiles for large scale lncRNA–miRNA interaction profiling**. 2017, **34**(5):812-819.
13. Volders P-J, Helsens K, Wang X, Menten B, Martens L, Gevaert K, Vandesompele J, Mestdagh P: **LNCipedia: a database for annotated human lncRNA transcript sequences and structures**. *Nucleic acids research* 2013, **41**(D1):D246-D251.
14. Betel D, Wilson M, Gabow A, Marks DS, Sander C: **The microRNA. org resource: targets and expression**. *Nucleic acids research* 2008, **36**(suppl 1):D149-D153.
15. Chen X, Yin J, Qu J, Huang LJPcb: **MDHGI: Matrix Decomposition and Heterogeneous Graph Inference for miRNA-disease association prediction**. 2018, **14**(8):e1006418.
16. Yoon J-H, Abdelmohsen K, Gorospe M: **Functional interactions among microRNAs and long noncoding RNAs**. In: *Seminars in cell & developmental biology: 2014*. Elsevier: 9-14.
17. Huang Z, Shi J, Gao Y, Cui C, Zhang S, Li J, Zhou Y, Cui QJNar: **HMDD v3. 0: a database for experimentally supported human microRNA–disease associations**. 2018, **47**(D1):D1013-D1017.
18. 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.

19. Xie B, Ding Q, Han H, Wu DJB: **miRCancer: a microRNA–cancer association database constructed by text mining on literature**. 2013, **29**(5):638-644.
20. Jiang Q, Wang Y, Hao Y, Juan L, Teng M, Zhang X, Li M, Wang G, Liu Y: **miR2Disease: a manually curated database for microRNA deregulation in human disease**. *Nucleic Acids Res* 2009, **37**(Database issue):D98-104.
21. 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* 2019, **47**(D1):D1013-D1017.
22. Yang Z, Ren F, Liu C, He S, Sun G, Gao Q, Yao L, Zhang Y, Miao R, Cao Y: **dbDEMC: a database of differentially expressed miRNAs in human cancers**. In: *BMC genomics: 2010*. Springer: S5.
23. Wang D, Gu J, Wang T, Ding Z: **OncomiRDB: a database for the experimentally verified oncogenic and tumor-suppressive microRNAs**. *Bioinformatics* 2014, **30**(15):2237-2238.
24. Khurana R, Verma VK, Rawoof A, Tiwari S, Nair RA, Mahidhara G, Idris MM, Clarke AR, Kumar LD: **OncomiRdbB: a comprehensive database of microRNAs and their targets in breast cancer**. *Bmc Bioinformatics* 2014, **15**(1):15.
25. Chen X, Xie D, Zhao Q, You ZH: **MicroRNAs and complex diseases: from experimental results to computational models**. *Brief Bioinform* 2019, **20**(2):515-539.
26. Kozomara A, Griffiths-Jones S: **miRBase: annotating high confidence microRNAs using deep sequencing data**. *Nucleic acids research* 2014, **42**(D1):D68-D73.
27. Nam S, Kim B, Shin S, Lee S: **miRGator: an integrated system for functional annotation of microRNAs**. *Nucleic Acids Res* 2008, **36**(Database issue):D159-164.
28. Megraw M, Sethupathy P, Corda B, Hatzigeorgiou AG: **miRGen: a database for the study of animal microRNA genomic organization and function**. *Nucleic Acids Res* 2007, **35**(Database issue):D149-155.
29. Girijadevi R, Sreedevi VC, Sreedharan JV, Pillai MR: **IntmiR: a complete catalogue of intronic miRNAs of human and mouse**. *Bioinformation* 2011, **5**(10):458-459.
30. Barh D, Bhat D, Viero C: **miReg: a resource for microRNA regulation**. *J Integr Bioinform* 2010, **7**(1):55-62.
31. Xiao F, Zuo Z, Cai G, Kang S, Gao X, Li TJNar: **miRecords: an integrated resource for microRNA–target interactions**. 2009, **37**(suppl_1):D105-D110.
32. Dweep H, Sticht C, Pandey P, Gretz NJJobi: **miRWalk–database: prediction of possible miRNA binding sites by "walking" the genes of three genomes**. 2011, **44**(5):839-847.
33. Rukov JL, Shomron N: **MicroRNA pharmacogenomics: post-transcriptional regulation of drug response**. *Trends in molecular medicine* 2011, **17**(8):412-423.
34. Shah MY, Ferrajoli A, Sood AK, Lopez-Berestein G, Calin GA: **microRNA therapeutics in cancer—an emerging concept**. *EBioMedicine* 2016, **12**:34-42.
35. Zheng C-H, Huang D-S, Zhang L, Kong X-ZJITotITiB: **Tumor clustering using nonnegative matrix factorization with gene selection**. 2009, **13**(4):599-607.
36. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z *et al*: **DrugBank 5.0: a major update to the DrugBank database for 2018**. *Nucleic Acids Res* 2018, **46**(D1):D1074-D1082.

37. Song J, Ye A, Jiang E, Yin X, Chen Z, Bai G, Zhou Y, Liu JJJocb: **Reconstruction and analysis of the aberrant lncRNA-miRNA-mRNA network based on competitive endogenous RNA in CESC**. 2018, **119**(8):6665-6673.
38. Liu J, Li H, Zheng B, Sun L, Yuan Y, Xing CJDd, sciences: **Competitive Endogenous RNA (ceRNA) Regulation Network of lncRNA–miRNA–mRNA in Colorectal Carcinogenesis**. 2019:1-10.
39. Chen X: **KATZLDA: KATZ measure for the lncRNA-disease association prediction**. *Scientific reports* 2015, **5**:16840.
40. Ma Y, Ge L, Ma Y, Jiang X, He T, Hu X: **Kernel Soft-neighborhood Network Fusion for MiRNA-Disease Interaction Prediction**. In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM): 2018*. IEEE: 197-200.
41. Qu Y, Zhang H, Liang C, Dong X: **Katzmda: prediction of mirna-disease associations based on katz model**. *IEEE Access* 2018, **6**:3943-3950.
42. Salmena L, Poliseno L, Tay Y, Kats L, Pandolfi PPJC: **A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language?** 2011, **146**(3):353-358.
43. Shi Y, Yang F, Wei S, Xu G: **Identification of Key Genes Affecting Results of Hyperthermia in Osteosarcoma Based on Integrative ChIP-Seq/TargetScan Analysis**. *Med Sci Monit* 2017, **23**:2042-2048.
44. Huang YA, Hu PW, Chan KCC, You ZH: **Graph convolution for predicting associations between miRNA and drug resistance**. *Bioinformatics* 2020, **36**(3):851-858.
45. Hu P, Huang YA, Chan KCC, You ZH: **Learning Multimodal Networks from Heterogeneous Data for Prediction of lncRNA-miRNA Interactions**. *IEEE/ACM Trans Comput Biol Bioinform* 2019.
46. Newman ME: **Clustering and preferential attachment in growing networks**. *Phys Rev E Stat Nonlin Soft Matter Phys* 2001, **64**(2 Pt 2):025102.
47. Crnic J: **Introduction to Modern Information Retrieval**. *Library Management* 2011, **32**(4-5):373-+.
48. Barabasi AL, Jeong H, Neda Z, Ravasz E, Schubert A, Vicsek T: **Evolution of the social network of scientific collaborations**. *Physica A* 2002, **311**(3-4):590-614.
49. Newman ME: **The structure of scientific collaboration networks**. *Proc Natl Acad Sci U S A* 2001, **98**(2):404-409.
50. Katz LJP: **A new status index derived from sociometric analysis**. 1953, **18**(1):39-43.
51. Papadimitriou A, Symeonidis P, Manolopoulos Y: **Friendlink: link prediction in social networks via bounded local path traversal**. In: *2011 International Conference on Computational Aspects of Social Networks (CASoN): 2011*. IEEE: 66-71.
52. Lu LY, Jin CH, Zhou T: **Similarity index based on local paths for link prediction of complex networks**. *Physical Review E* 2009, **80**(4):046122.
53. Chen H-H, Gou L, Zhang X, Giles CL: **Discovering missing links in networks using vertex similarity measures**. In: *Proceedings of the 27th annual ACM symposium on applied computing: 2012*. 138-143.
54. Lichtenwalter RN, Chawla NV: **Vertex collocation profiles: subgraph counting for link analysis and prediction**. In: *Proceedings of the 21st international conference on World Wide Web: 2012*. 1019-1028.
55. Spitzer F: **Principles of random walk**, vol. 34: Springer Science & Business Media; 2013.

56. Mei Q, Zhou D, Church K: **Query suggestion using hitting time**. In: *Proceedings of the 17th ACM conference on Information and knowledge management: 2008*. 469-478.
57. Jeh G, Widom J: **SimRank: a measure of structural-context similarity**. In: *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining: 2002*. 538-543.
58. Yan S, Xu D, Zhang B, Zhang H-J, Yang Q, Lin SJItopa, intelligence m: **Graph embedding and extensions: A general framework for dimensionality reduction**. 2006, **29**(1):40-51.
59. Cai D, He X, Han J: **Spectral regression: a unified subspace learning framework for content-based image retrieval**. In: *Proceedings of the 15th ACM international conference on Multimedia: 2007*. 403-412.
60. He X, Niyogi P: **Locality preserving projections**. In: *Advances in neural information processing systems: 2004*. 153-160.
61. Gong C, Tao D, Yang J, Fu K: **Signed laplacian embedding for supervised dimension reduction**. In: *Twenty-Eighth AAAI Conference on Artificial Intelligence: 2014*.
62. Hofmann T, Buhmann J: **Multidimensional scaling and data clustering**. In: *Advances in neural information processing systems: 1995*. 459-466.
63. Roweis ST, Saul LK: **Nonlinear dimensionality reduction by locally linear embedding**. *Science* 2000, **290**(5500):2323-2326.
64. Jiang R, Fu WJ, Wen L, Hao SJ, Hong RC: **Dimensionality reduction on Anchorgraph with an efficient Locality Preserving Projection**. *Neurocomputing* 2016, **187**:109-118.
65. Yang Y, Nie F, Xiang S, Zhuang Y, Wang W: **Local and global regressive mapping for manifold learning with out-of-sample extrapolation**. In: *Twenty-Fourth AAAI Conference on Artificial Intelligence: 2010*.
66. Xiang S, Nie F, Zhang C, Zhang CJIToK, Engineering D: **Nonlinear dimensionality reduction with local spline embedding**. 2008, **21**(9):1285-1298.
67. Allab K, Labiod L, Nadif M: **A Semi-NMF-PCA Unified Framework for Data Clustering**. *Ieee T Knowl Data En* 2017, **29**(1):2-16.
68. Mosheyev L, Michael ZJOM, Software: **Penalty/Barrier multiplier algorthm for semidefinit programming***. 2000, **13**(4):235-261.
69. Cao S, Wei L, Xu Q: **GraRep: Learning Graph Representations with Global Structural Information**. In*: 2015*.
70. Golub GH, Reinsch C: **Singular Value Decomposition and Least Squares Solutions**. *Numerische Mathematik* 1970, **14**(5):403-&.
71. Ou M, Peng C, Jian P, Zhang Z, Zhu W: **Asymmetric Transitivity Preserving Graph Embedding**. In: *the 22nd ACM SIGKDD International Conference: 2016*.
72. Ahmed A, Shervashidze N, Narayanamurthy S, Josifovski V, Smola AJ: **Distributed Large-scale Natural Graph Factorization**. In: *Proceedings of the 22nd international conference on World Wide Web: 2013*.
73. Sun G, Zhang X: **Graph Embedding with Rich Information through Heterogeneous Network**.
74. Yang C, Liu ZJCS: **Comprehend DeepWalk as Matrix Factorization**. 2015.
75. Cheng W, Greaves C, Linguistics MWJlJoC: **From n-gram to skipgram to concgram**. **11**(4):411-433.
76. Hochreiter S, Schmidhuber J: **Long short-term memory**. *Neural Comput* 1997,

9(8):1735-1780.

77. Zhao XH, Chang A, Das Sarma A, Zheng HT, Zhao BY: **On the Embeddability of Random Walk Distances**. *Proceedings of the Vldb Endowment* 2013, **6**(14):1690-1701.
78. Wu F, Lu X, Song J, Yan S, Zhang ZM, Rui Y, Zhuang Y: **Learning of Multimodal Representations With Random Walks on the Click Graph**. *IEEE Trans Image Process* 2016, **25**(2):630-642.
79. Wang C, Pan S, Long G, Zhu X, Jing J: **MGAE: Marginalized Graph Autoencoder for Graph Clustering**. In: *the 2017 ACM: 2017*.
80. Wang D, Peng C, Zhu W: **Structural Deep Network Embedding**. In: *the 22nd ACM SIGKDD International Conference: 2016*.
81. Cao S: **deep neural network for learning graph representations**. In: *AAAI: 2016*.
82. Tian F, Gao B, Cui Q, Chen E, Liu T: **Learning deep representations for graph clustering**. In: *national conference on artificial intelligence: 2014*. 1293-1299.
83. Niepert M, Ahmed M, Kutzkov K: **Learning convolutional neural networks for graphs**. In: *International conference on machine learning: 2016*. 2014-2023.
84. Henaff M, Bruna J, Lecun Y: **Deep Convolutional Networks on Graph-Structured Data**.
85. Defferrard M, Bresson X, Vandergheynst P: **Convolutional neural networks on graphs with fast localized spectral filtering**. In: *Advances in Neural Information Processing Systems: 2016*. 3844-3852.
86. Kipf TN, Welling M: **Semi-supervised classification with graph convolutional networks**. *arXiv preprint arXiv:160902907* 2016.
87. Chen X, Yan CC, Zhang X, Li Z, Deng L, Zhang Y, Dai Q: **RBMMMDA: predicting multiple types of disease-microRNA associations**. *Sci Rep* 2015, **5**(1):13877.
88. Chen X, Yan CC, Zhang X, You ZH, Huang YA, Yan GY: **HGIMDA: Heterogeneous graph inference for miRNA-disease association prediction**. *Oncotarget* 2016, **7**(40):65257-65269.
89. Jiang QH, Hao YY, Wang GH, Juan LR, Zhang TJ, Teng MX, Liu YL, Wang YD: **Prioritization of disease microRNAs through a human phenome-microRNAome network**. *Bmc Systems Biology* 2010, **4**:1-9.
90. Sun D, Li A, Feng H, Wang M: **NTSMDA: prediction of miRNA-disease associations by integrating network topological similarity**. *Mol Biosyst* 2016, **12**(7):2224-2232.
91. Chen X, Wu QF, Yan GY: **RKNNMDA: Ranking-based KNN for MiRNA-Disease Association prediction**. *RNA Biol* 2017, **14**(7):952-962.
92. Chen X, Yan GY: **Semi-supervised learning for potential human microRNA-disease associations inference**. *Sci Rep* 2014, **4**:5501.
93. Xuan P, Han K, Guo Y, Li J, Li X, Zhong Y, Zhang Z, Ding J: **Prediction of potential disease-associated microRNAs based on random walk**. *Bioinformatics* 2015, **31**(11):1805-1815.
94. Luo J, Huang C, Ding P: **A Meta-Path-Based Prediction Method for Human miRNA-Target Association**. *Biomed Res Int* 2016, **2016**:7460740.
95. You ZH, Huang ZA, Zhu Z, Yan GY, Li ZW, Wen Z, Chen X: **PBMDA: A novel and effective path-based computational model for miRNA-disease association prediction**. *PLoS Comput Biol* 2017, **13**(3):e1005455.
96. Xu C, Ping Y, Li X, Zhao H, Wang L, Fan H, Xiao Y, Li X: **Prioritizing candidate disease miRNAs by integrating phenotype associations of multiple diseases with matched**

miRNA and mRNA expression profiles. *Mol Biosyst* 2014, **10**(11):2800-2809.

97. Pasquier C, Gardes J: **Prediction of miRNA-disease associations with a vector space model**. *Sci Rep* 2016, **6**(1):27036.
98. Li JQ, Rong ZH, Chen X, Yan GY, You ZH: **MCMDA: Matrix completion for MiRNA-disease association prediction**. *Oncotarget* 2017, **8**(13):21187-21199.
99. Parikshak NN, Swarup V, Belgard TG, Irimia M, Ramaswami G, Gandal MJ, Hartl C, Leppa V, Ubieta LT, Huang J *et al*: **Genome-wide changes in lncRNA, splicing, and regional gene expression patterns in autism**. *Nature* 2016, **540**(7633):423-427.
100. Salmena L, Poliseno L, Tay Y, Kats L, Pandolfi PP: **A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language?** *Cell* 2011, **146**(3):353-358.
101. Zhang GL, Pian C, Chen Z, Zhang J, Xu MM, Zhang LY, Chen YY: **Identification of cancer-related miRNA-lncRNA biomarkers using a basic miRNA-lncRNA network**. *Plos One* 2018, **13**(5):e0196681.
102. Yuan W, Li X, Liu L, Wei C, Sun D, Peng S, Jiang L: **Comprehensive analysis of lncRNA-associated ceRNA network in colorectal cancer**. *Biochem Biophys Res Commun* 2019, **508**(2):374-379.
103. Lopes CT, Franz M, Kazi F, Donaldson SL, Morris Q, Bader GD: **Cytoscape Web: an interactive web-based network browser**. *Bioinformatics* 2010, **26**(18):2347-2348.
104. Li J, Han X, Wan Y, Zhang S, Zhao Y, Fan R, Cui Q, Zhou Y: **TAM 2.0: tool for MicroRNA set analysis**. *Nucleic Acids Res* 2018, **46**(W1):W180-W185.
105. Pinzon N, Li B, Martinez L, Sergeeva A, Presumey J, Apparailly F, Seitz H: **microRNA target prediction programs predict many false positives**. *Genome Research* 2017, **27**(2):234-245.
106. Chen X, Xie D, Zhao Q, You Z-HJBib: **MicroRNAs and complex diseases: from experimental results to computational models**. 2017, **20**(2):515-539.
107. Ning S, Yue M, Wang P, Liu Y, Zhi H, Zhang Y, Zhang J, Gao Y, Guo M, Zhou DJNar: **LincSNP 2.0: an updated database for linking disease-associated SNPs to human long non-coding RNAs and their TFBSs**. 2016:gkw945.
108. He X, Chen T, Kan M-Y, Chen X: **Trirank: Review-aware explainable recommendation by modeling aspects**. In: *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management: 2015*. ACM: 1661-1670.
109. Chen X, Yin J, Qu J, Huang L: **MDHGI: Matrix Decomposition and Heterogeneous Graph Inference for miRNA-disease association prediction**. *PLoS Comput Biol* 2018, **14**(8):e1006418.
110. Zeng XX, Liu L, Lu LY, Zou Q: **Prediction of potential disease-associated microRNAs using structural perturbation method**. *Bioinformatics* 2018, **34**(14):2425-2432.
111. Jiang L, Ding Y, Tang J, Guo F: **MDA-SKF: Similarity Kernel Fusion for Accurately Discovering miRNA-Disease Association**. *Front Genet* 2018, **9**:618.
112. Sarwar BM, Karypis G, Konstan JA, Riedl JJW: **Item-based collaborative filtering recommendation algorithms**. 2001, **1**:285-295.
113. Koren Y: **Factorization meets the neighborhood: a multifaceted collaborative filtering model**. In: *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining: 2008*. ACM: 426-434.
114. Chen X, Huang Y-A, You Z-H, Yan G-Y, Wang X-SJB: **A novel approach based on KATZ**

measure to predict associations of human microbiota with non-infectious diseases. 2016, **33**(5):733-739.

115. Yang Y, Fu X, Qu W, Xiao Y, Shen H-BJB: **MiRGOFS: A GO-based functional similarity measure for miRNAs, with applications to the prediction of miRNA subcellular localization and miRNA-disease association**. 2018.
116. Cheng L, Hu Y, Sun J, Zhou M, Jiang Q: **DincRNA: a comprehensive web-based bioinformatics toolkit for exploring disease associations and ncRNA function**. *Bioinformatics* 2018, **34**(11):1953-1956.
117. Hafner M, Niepel M, Sorger PK: **Alternative drug sensitivity metrics improve preclinical cancer pharmacogenomics**. *Nature biotechnology* 2017, **35**(6):500.
118. Hopkins AL, Groom CR: **The druggable genome**. *Nature reviews Drug discovery* 2002, **1**(9):727.
119. Overington JP, Al-Lazikani B, Hopkins AL: **How many drug targets are there?** *Nature reviews Drug discovery* 2006, **5**(12):993.
120. Matsui M, Corey DR: **Non-coding RNAs as drug targets**. *Nature reviews Drug discovery* 2017, **16**(3):167.
121. Chavali PL, Funa K, Chavali S: **Cis-regulation of microRNA expression by scaffold/matrix-attachment regions**. *Nucleic Acids Res* 2011, **39**(16):6908-6918.
122. Rupaimoole R, Slack FJ: **MicroRNA therapeutics: towards a new era for the management of cancer and other diseases**. *Nature reviews Drug discovery* 2017, **16**(3):203.
123. Roberti A, Sala DL, Cinti C: **Multiple genetic and epigenetic interacting mechanisms contribute to clonally selection of drug-resistant tumors: Current views and new therapeutic prospective**. *Journal of cellular physiology* 2006, **207**(3):571-581.
124. Lehnert M: **Chemotherapy resistance in breast cancer**. *Anticancer research* 1998, **18**(3C):2225-2226.
125. Bolton EE, Wang Y, Thiessen PA, Bryant SH: **PubChem: integrated platform of small molecules and biological activities**. In: *Annual reports in computational chemistry.* vol. 4: Elsevier; 2008: 217-241.
126. Shrive FM, Stuart H, Quan H, Ghali WA: **Dealing with missing data in a multi-question depression scale: a comparison of imputation methods**. *BMC Med Res Methodol* 2006, **6**(1):57.
127. Yang Y, Fu X, Qu W, Xiao Y, Shen H-B: **MiRGOFS: A GO-based functional similarity measure for miRNAs, with applications to the prediction of miRNA subcellular localization and miRNA-disease association**. *Bioinformatics* 2018.
128. Chavali S, Bruhn S, Tiemann K, Saetrom P, Barrenas F, Saito T, Kanduri K, Wang H, Benson M: **MicroRNAs act complementarily to regulate disease-related mRNA modules in human diseases**. *RNA* 2013, **19**(11):1552-1562.
129. Anokye-Danso F, Trivedi CM, Juhr D, Gupta M, Cui Z, Tian Y, Zhang Y, Yang W, Gruber PJ, Epstein JA *et al*: **Highly efficient miRNA-mediated reprogramming of mouse and human somatic cells to pluripotency**. *Cell Stem Cell* 2011, **8**(4):376-388.
130. Wang T, Li L, Huang YA, Zhang H, Ma Y, Zhou X: **Prediction of Protein-Protein Interactions from Amino Acid Sequences Based on Continuous and Discrete Wavelet Transform Features**. *Molecules* 2018, **23**(4):823.

131. Atwood J, Towsley D: **Diffusion-convolutional neural networks**. In: *Advances in Neural Information Processing Systems: 2016*. 1993-2001.
132. Duvenaud DK, Maclaurin D, Iparraguirre J, Bombarell R, Hirzel T, Aspuru-Guzik A, Adams RP: **Convolutional networks on graphs for learning molecular fingerprints**. In: *Advances in neural information processing systems: 2015*. 2224-2232.
133. Glorot X, Bengio Y: **Understanding the difficulty of training deep feedforward neural networks**. In: *Proceedings of the thirteenth international conference on artificial intelligence and statistics: 2010*. 249-256.
134. Ge M, Li A, Wang M: **A bipartite network-based method for prediction of long non-coding RNA–protein interactions**. *Genomics, proteomics & bioinformatics* 2016, **14**(1):62-71.
135. Chen X, Huang Y-A, You Z-H, Yan G-Y, Wang X-S: **A novel approach based on KATZ measure to predict associations of human microbiota with non-infectious diseases**. *Bioinformatics* 2016, **33**(5):733-739.
136. Huang Y-A, Chan KC, You Z-H: **Constructing Prediction Models from Expression Profiles for Large Scale lncRNA-miRNA Interaction Profiling**. *Bioinformatics* 2017.
137. Su X, Khoshgoftaar TM: **A survey of collaborative filtering techniques**. *Advances in artificial intelligence* 2009, **2009**:4.
138. Boutsidis C, Gallopoulos E: **SVD based initialization: A head start for nonnegative matrix factorization**. *Pattern Recognition* 2008, **41**(4):1350-1362.
139. Lin D: **An information-theoretic definition of similarity**. In: *Icml: 1998*. Citeseer: 296-304.
140. Resnik P: **Using information content to evaluate semantic similarity in a taxonomy**. *arXiv preprint cmp-lg/9511007* 1995.
141. Sqalli MH, Al-Saeedi M, Binbeshr F, Siddiqui M: **UCloud: A simulated Hybrid Cloud for a university environment**. In: *Cloud Networking (CLOUDNET), 2012 IEEE 1st International Conference on: 2012*. IEEE: 170-172.
142. Cheng L, Li J, Ju P, Peng J, Wang Y: **SemFunSim: a new method for measuring disease similarity by integrating semantic and gene functional association**. *PloS one* 2014, **9**(6):e99415.
143. Zhu X, Goldberg AB: **Introduction to semi-supervised learning**. *Synthesis lectures on artificial intelligence and machine learning* 2009, **3**(1):1-130.
144. Salmena L, Poliseno L, Tay Y, Kats L, Pandolfi PP: **A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language?** *Cell* 2011, **146**(3):353-358.
145. Liu B, Fang L, Liu F, Wang X, Chou K-C: **iMiRNA-PseDPC: microRNA precursor identification with a pseudo distance-pair composition approach**. *Journal of Biomolecular Structure and Dynamics* 2016, **34**(1):223-235.
146. Liu B, Fang L, Liu F, Wang X, Chen J, Chou K-C: **Identification of real microRNA precursors with a pseudo structure status composition approach**. *PloS one* 2015, **10**(3):e0121501.
147. Liu B, Liu F, Fang L, Wang X, Chou K-C: **repRNA: a web server for generating various feature vectors of RNA sequences**. *Molecular Genetics and Genomics* 2016, **291**(1):473-481.
148. Liu B, Liu F, Wang X, Chen J, Fang L, Chou K-C: **Pse-in-One: a web server for generating**

various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic acids research* 2015, **43**(W1):W65-W71.

149. Quinn JJ, Chang HY: **Unique features of long non-coding RNA biogenesis and function**. *Nature Reviews Genetics* 2016, **17**(1):47-62.
150. Li J-H, Liu S, Zhou H, Qu L-H, Yang J-H: **starBase v2. 0: decoding miRNA-ceRNA, miRNA-ncRNA and protein–RNA interaction networks from large-scale CLIP-Seq data**. *Nucleic acids research* 2013:gkt1248.
151. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL: **Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses**. *Genes & development* 2011, **25**(18):1915-1927.
152. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG: **The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression**. *Genome research* 2012, **22**(9):1775-1789.
153. Xia T, Liao Q, Jiang X, Shao Y, Xiao B, Xi Y, Guo J: **Long noncoding RNA associated-competing endogenous RNAs in gastric cancer**. *Scientific reports* 2014, **4**:6088.
154. Ballantyne MD, McDonald RA, Baker AH: **lncRNA/MicroRNA interactions in the vasculature**. *Clinical Pharmacology & Therapeutics* 2016, **99**(5):494-501.
155. Du Z, Sun T, Hacisuleyman E, Fei T, Wang X, Brown M, Rinn JL, Lee MG-S, Chen Y, Kantoff PW: **Integrative analyses reveal a long noncoding RNA-mediated sponge regulatory network in prostate cancer**. *Nature communications* 2016, **7**.
156. Poliseno L, Pandolfi PP: **PTEN ceRNA networks in human cancer**. *Methods* 2015, **77**:41-50.
157. Li J, Ma W, Zeng P, Wang J, Geng B, Yang J, Cui Q: **LncTar: a tool for predicting the RNA targets of long noncoding RNAs**. *Briefings in bioinformatics* 2015, **16**(5):806-812.
158. Cesana M, Daley GQ: **Deciphering the rules of ceRNA networks**. *Proceedings of the National Academy of Sciences* 2013, **110**(18):7112-7113.
159. Ala U, Karreth FA, Bosia C, Pagnani A, Taulli R, Léopold V, Tay Y, Provero P, Zecchina R, Pandolfi PP: **Integrated transcriptional and competitive endogenous RNA networks are cross-regulated in permissive molecular environments**. *Proceedings of the National Academy of Sciences* 2013, **110**(18):7154-7159.
160. Buchler NE, Louis M: **Molecular titration and ultrasensitivity in regulatory networks**. *Journal of molecular biology* 2008, **384**(5):1106-1119.
161. Levine E, Hwa T: **Small RNAs establish gene expression thresholds**. *Current opinion in microbiology* 2008, **11**(6):574-579.
162. Mukherji S, Ebert MS, Zheng GX, Tsang JS, Sharp PA, van Oudenaarden A: **MicroRNAs can generate thresholds in target gene expression**. *Nature genetics* 2011, **43**(9):854-859.
163. Gong J, Liu W, Zhang J, Miao X, Guo AY: **lncRNASNP: a database of SNPs in lncRNAs and their potential functions in human and mouse**. *Nucleic Acids Res* 2015, **43**(Database issue):D181-186.
164. Hsu S-D, Lin F-M, Wu W-Y, Liang C, Huang W-C, Chan W-L, Tsai W-T, Chen G-Z, Lee C-J, Chiu C-M: **miRTarBase: a database curates experimentally validated microRNA–target interactions**. *Nucleic acids research* 2010:gkq1107.

165. Bu D, Yu K, Sun S, Xie C, Skogerbø G, Miao R, Xiao H, Liao Q, Luo H, Zhao G: **NONCODE v3. 0: integrative annotation of long noncoding RNAs**. *Nucleic acids research* 2011:gkr1175.
166. Yang S, Ning Q, Zhang G, Sun H, Wang Z, Li Y: **Construction of differential mRNA-lncRNA crosstalk networks based on ceRNA hypothesis uncover key roles of lncRNAs implicated in esophageal squamous cell carcinoma**. *Oncotarget* 2016, **7**(52):85728.
167. Li Y, Chen J, Zhang J, Wang Z, Shao T, Jiang C, Xu J, Li X: **Construction and analysis of lncRNA-lncRNA synergistic networks to reveal clinically relevant lncRNAs in cancer**. *Oncotarget* 2015, **6**(28):25003.
168. Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B: **Biopython: freely available Python tools for computational molecular biology and bioinformatics**. *Bioinformatics* 2009, **25**(11):1422-1423.
169. Deng S-P, Zhu L, Huang D-SJIAToCB, Bioinformatics: **Predicting hub genes associated with cervical cancer through gene co-expression networks**. 2016, **13**(1):27-35.
170. Zhu L, Deng S-P, Huang D-SJIton: **A two-stage geometric method for pruning unreliable links in protein-protein networks**. 2015, **14**(5):528-534.
171. You Z-H, Yin Z, Han K, Huang D-S, Zhou XJBB: **A semi-supervised learning approach to predict synthetic genetic interactions by combining functional and topological properties of functional gene network**. 2010, **11**(1):343.
172. Huang D-S, Zheng C-HJB: **Independent component analysis-based penalized discriminant method for tumor classification using gene expression data**. 2006, **22**(15):1855-1862.
173. Huang D-S, Huang XJP, letters p: **Improved performance in protein secondary structure prediction by combining multiple predictions**. 2006, **13**(10):985-991.
174. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, Haibe-Kains B, Goldenberg A: **Similarity network fusion for aggregating data types on a genomic scale**. *Nat Methods* 2014, **11**(3):333-337.
175. Kornienko AE, Dotter CP, Guenzl PM, Gisslinger H, Gisslinger B, Cleary C, Kralovics R, Pauler FM, Barlow DP: **Long non-coding RNAs display higher natural expression variation than protein-coding genes in healthy humans**. *Genome biology* 2016, **17**(1):14.
176. Zhao Y, Li H, Fang S, Kang Y, Hao Y, Li Z, Bu D, Sun N, Zhang MQ, Chen R: **NONCODE 2016: an informative and valuable data source of long non-coding RNAs**. *Nucleic acids research* 2016, **44**(D1):D203-D208.
177. Yuan L, Zhu L, Guo W-L, Zhou X, Zhang Y, Huang Z, Huang D-SJIAToCB, Bioinformatics: **Nonconvex penalty based low-rank representation and sparse regression for eQTL mapping**. 2017, **14**(5):1154-1164.