Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 7460740, 9 pages
http://dx.doi.org/10.1155/2016/7460740

*Research Article*

# A Meta-Path-Based Prediction Method for Human miRNA-Target Association

**Jiawei Luo,[1] Cong Huang,[2] and Pingjian Ding[2]**

[1] *College of Information Science and Electronic Engineering & Collaboration and Innovation Center for Digital Chinese Medicine of 2011 Project of Colleges and Universities in Hunan Province, Hunan University, Changsha, Hunan 410082, China*
[2] *College of Information Science and Electronic Engineering, Hunan University, Changsha, Hunan 410082, China*

Correspondence should be addressed to Jiawei Luo; luojiawei@hnu.edu.cn

Received 30 June 2016; Revised 14 August 2016; Accepted 21 August 2016

Academic Editor: Xing Chen

MicroRNAs (miRNAs) are short noncoding RNAs that play important roles in regulating gene expressing, and the perturbed miRNAs are often associated with development and tumorigenesis as they have effects on their target mRNA. Predicting potential miRNA-target associations from multiple types of genomic data is a considerable problem in the bioinformatics research. However, most of the existing methods did not fully use the experimentally validated miRNA-mRNA interactions. Here, we developed RMLM and RMLMSe to predict the relationship between miRNAs and their targets. RMLM and RMLMSe are global approaches as they can reconstruct the missing associations for all the miRNA-target simultaneously and RMLMSe demonstrates that the integration of sequence information can improve the performance of RMLM. In RMLM, we use RM measure to evaluate different relatedness between miRNA and its target based on different meta-paths; logistic regression and MLE method are employed to estimate the weight of different meta-paths. In RMLMSe, sequence information is utilized to improve the performance of RMLM. Here, we carry on fivefold cross validation and pathway enrichment analysis to prove the performance of our methods. The fivefold experiments show that our methods have higher AUC scores compared with other methods and the integration of sequence information can improve the performance of miRNA-target association prediction.

## 1. Introduction

MicroRNAs (miRNAs) are important endogenous 21-22 nt RNAs that play important regulatory roles in gene expression. Several studies have shown that miRNAs participate in the regulation of amount cellular process, such as cell proliferation and differentiation [1], development [2], and disease [3, 4]. Considering the importance of miRNAs, it is critical to identify and decipher miRNA-target interactions at a genome level.

All the time, scientists and academics have made great efforts in uncovering the associations between miRNA and its targets by using biological experiments [5–8]. However, it is impossible to depict a complete picture of miRNA regulation mechanisms only relying on biological experiments due to the high expenses on time and cost [9]. Therefore, computational approaches must be designed to be a cost-effective choice to describe the complete mechanism of miRNA regulatory. Now, many computational approaches show great advantage in predicting putative miRNA targets [10–13].

Over the past decade, plenty of miRNA-mRNA pairs prediction approaches have been developed to identify miRNA targets by using sequence data, including TargetScanS/TargetScan [14, 15], miRanda [16], Pictar [17], DITAT-MicroT [18], and PITA [19]. The majority of these prediction algorithms were built on specific binding rules, including the degree of site conservation, thermodynamic stability, sequence complementarity, energy, target site context, secondary structure, and site accessibility. Because of the complex character of miRNA-target interactions, these sequence-based methods have relatively high false-positive rate [20]. Furthermore, those predictions methods were mostly only at static sequence level, leading to those exact interactions that are specific to certain conditions or diseases. More importantly, sequence-based methods do not support statistically significant predictions as the miRNA binding sites are small, causing the results by different methods to be inconsistent.

To identify condition-specific interactions, many methods integrating expression profiles information into sequence-based predictions have been proposed to study miRNA-mRNA regulatory mechanism. These methods are based on the assumption that gene has negative correlations with the miRNA because of the downregulation effect that miRNAs have on their targets. These methods can be divided into four categories including simple correlation analysis [21, 22], simple/regularized regression models [23–25], Bayesian inference [19, 26], and causally inference between miRNAs and their targets [27]. Pearson correlation, one of the typical simple correlation methods, is commonly used in computing the strength of the association between a pair of miRNA and mRNA. However, Pearson correlation has high false-positive rate as the simplicity of it. Furthermore, Pearson correlation is mainly used in predicting linear associations. Lasso regression [24, 25], one of the regression models, is a high-dimensional method used to extract more reliable association as they usually optimize the network provided by sequence-based method and retain the relatively reliable edges. GenMir++ [19], the first and well-cited Bayesian inference method, calculates the existence probabilities of the relationship between a miRNA and its target based on a Bayesian model. However, this method needs prior information, such as sequence information. In general, methods in Bayesian category assume different priors [28] and are difficult in learning parameters. MCMG (joint analysis of multiple cancer for MiRNA-gene interactions), based on empirical Bayesian model [29], identifies miRNA-target associations that are either specific to a cancer type or common to several cancers by jointly analyzed across cancers. Muniategui et al. use do-calculus to estimate the causal effects the miRNA have on all the target mRNAs. The four categories methods can improve prediction performance as they integrate expression profiles information into sequence-based prediction methods [30]. But, most of the existing approaches cannot effectively use the valuable experimentally validated information [31–34]. Besides, the lack of miRNA expression profile may cause the unreliability of the predicted miRNA-target associations.

On the whole, the limitations of existing methods are summarized as follows. Firstly, sequenced-based prediction algorithms suffer from a high false-positive rate; second, the methods integrating expression profile data can only analyse one cancer every time; third, some methods cannot effectively utilize validated knowledge. To solve these problems, we propose two network-based approaches, RMLM and RMLMSe, to identify miRNA-target interactions based on meta-path. Meta-path is a good measuring method to compute the relatedness between the same or different types of objects in heterogeneous information network, as it contains a certain sequence of different link types [35]. Different meta-paths have different semantic meaning corresponding to different relationships between connected objects. In RMLM, we first utilize RM (a meta-path related measure proposed by Cao et al. [36]) to evaluate the existence probability of a link between miRNA and its targets. As different meta-path corresponds to different relation graphs, we may improve the final performance when integrating these different graphs by appropriate weights corresponding to different meta-paths. Thus, we then employ logistic regression and maximum-likelihood estimation (MLE) method to estimate the weight of different meta-path. Here, the issue of relationship prediction can be regarded as a two-class classification problem by using Bayesian analysis and logistic regression and then the MLE method can be employed to estimate the parameter vector. In RMLMSe, sequence information is integrated to improve the performance of the RMLM. Furthermore, as global approaches, RMLM and RMLMSe can remodel the missing relationship for all the diseases-associated miRNAs at the same time. Fivefold cross validations, pathway enrichment analysis about global network, and three important diseases network show that our proposed methods work well in predicting the relationship between miRNA and its target.

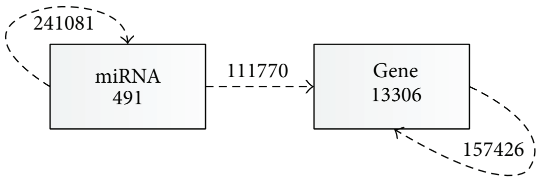

Figure 1: Network schema of the miRNA-target network. The network contains two types of objects, miRNA and its targets. Each box represents one type of nodes, and each dashed line represents one type of links. The numbers in the figure represent the numbers of nodes/links of different types.

## 2. Problem Definition

In this part, we describe the concepts of Heterogeneous Information Network and meta-path used in this paper.

*2.1. Heterogeneous Information Network.* A heterogeneous information network is an important type of information network with multiple types of nodes and multiple types of links [36–38]. It can be represented as $G = (V, E)$. $V$ is the set of nodes, which involves $n$ types of nodes: $V_1 = \{v_1^1, v_1^2, \ldots, v_1^x\}, \ldots, V_n = \{v_n^1, v_n^2, \ldots, v_n^y\}$, where $v_i^j$ is $j$th node of type $i$. $E \subseteq V \times V$ is the set of links between the nodes in $V$, which involves $m$ types of links.

Each type of links between source node of type $i$ and target node of type $j$ corresponds to a binary relation $R_{ij}$. More specifically, $R_{ij}^{st} = 1$ if $v_i^s$ ($s$th nodes of type $i$) and $v_j^t$ ($t$th nodes of type $j$) are connected by a link of type $R^{ij}$. For example, in Figure 1, the relation between miRNA and gene is "regulate." Particularly, $R_{ij}^{st}$ equals 1 if $s$th miRNA regulates $t$th gene.

Moreover, a weighted matrix $W_{ij} = |V_i| \times |V_j|$ can be used to describe the relation $R_{ij}$, where $W_{ij}^{st} \in [0, 1]$ is the existence probability of link between nodes $v_i^s$ and $v_j^t$. Particularly, $W_{ij}^{st} = 1$, if there exists an edge between $v_i^s$ and $v_j^t$. Otherwise, $W_{ij}^{st}$ is set as 0 in initialization for the unknown links.

*2.2. Meta-Path.* In heterogeneous information network, meta-path is defined on network schema. A meta-path $P$ is described in the form $A_1 \rightarrow A_2 \rightarrow \cdots A_{n-1} \rightarrow A_n$, where $A_i$ is $i$th type of object and a relation must exist from $A_{i-1}$ to $A_i$, $i = 2, 3, \ldots, n$. Similarly, we define the inverse path of $P$ as $P^{-1}$, denoted as $A_n \rightarrow A_{n-1} \rightarrow \cdots A_2 \rightarrow A_1$. Specifically, relation $\mathrm{A}_{i-1} \rightarrow A_i$ is the inverse relation of $A_i \rightarrow A_{i-1}$. For example, in Figure 1, a meta-path "gene → miRNA → gene" is a composite sequence between genes. The relation from miRNA to gene is "regulate" and the relation from gene to miRNA is "regulate$^{-1}$"; "regulate$^{-1}$" is the inverse relation of "regulate." Meta-path can connect object of the same or different types; thus, they can show knowledge between homologous objects or heterologous objects. For example, in Figure 1, for gene $i$ and gene $j$, they can connect through another gene $k$, gene $i$ → gene $k$ → gene $j$; this means gene $i$ and gene $j$ have relation with gene $k$ simultaneously and there may exist relation between gene $i$ and gene $j$ by information transfer. However, gene $i$ and gene $j$ can also connect by miRNA $k$, gene $i$ → miRNA $k$ → gene $j$; this means gene $i$ and gene $j$ are regulated by a common miRNA $k$ and there may exist relation between gene $i$ and gene $j$ by information transfer. Different meta-paths of different relations correspond to different relation graphs with different semantics. For example, in Figure 1, the meta-path "gene → gene" denotes that two genes are connected by "PPI" links, while the meta-path "gene → miRNA → gene" corresponds to the semantic that two genes are regulated by a common miRNA. Thus, similarity between the same or different type of nodes can be described by different meta-paths with different semantics.

In this paper, the meta-path from source node of type $i$ to target node of type $j$ is described as $P_{ij}$. Particularly, $P_{ii}$ is the meta-path between nodes of the same type $i$; $P_{iis}$ is $s$th meta-path of $P_{ii}$. $P_{jj}$ and $P_{jjt}$ are the same to $P_{ii}$ and $P_{iis}$. $P_{ijst}$ is a meta-path by connecting $P_{iis}$, $R_{ij}$, and $P_{jjt}$ in sequence; it can be written as a certain sequence of relations: $R_{k_0k_1}, R_{k_1k_2}, \ldots, R_{k_{n-1}k_n}$; here $k_0 = i,\ k_n = j$ and the length of $P_{ijst}$ is $n$.

## 3. Method

RMLM and RMLMSe consist of three steps. In the first step, we utilize MISIM (proposed by Wang et al. in [39]) to calculate the miRNA functional similarity matrix and then construct the heterogeneous network. Next, we calculate the relatedness between any miRNA and its targets and extract the feature vector of these interactions. In RMLM, the feature vector only contains different relatedness of different meta-path between miRNA and its targets. However, in RMLMSe, the feature vector not only contains different relatedness from different meta-path, but also contains feature extracted from sequence information. Finally, logistic regression and MLE method are employed to compute the different weights of different meta-paths. Sections 3.1–3.4 are the detailed introduction of RMLM. Section 3.5 is about RMLMSe.

### 3.1. Construction of the Heterogeneous Network

*3.1.1. miRNA-miRNA Similarity Estimation.* In [39], Wang et al. compute miRNA-miRNA functional similarity score based on the assumption that miRNAs with similar functions tend to be related to similar disease. To get the miRNA-miRNA similarity matrix, there contains three procedures. We take miRNA $i$ and miRNA $j$ as an example. First, we identify diseases that related to these two miRNAs, encoded as $D_i$ and $D_j$. We can obtain the relationship between miRNAs and diseases from The Human MicroRNA Disease Database (HMDD dataset). Then, we can calculate similarity of any pair of diseases using a hierarchical structure. The semantic similarity of disease is calculated based on directed acyclic graph obtained from the US National Library of Medicine in 2015 (MeSH, https://www.nlm.nih.gov/mesh/). Finally, we utilize the similarity score between $D_i$ and $D_j$ to compute the relatedness score between miRNA $i$ and miRNA $j$. In this paper, we use SM (a 491 × 491 matrix) to represent the miRNA-miRNA similarity matrix; $\mathrm{SM}(i, j)$ is the functional similarity score between miRNA $i$ and miRNA $j$.

*3.1.2. Construction of the Heterogeneous Network.* We construct the heterogeneous network by connecting the miRNA interaction network and PPI utilizing the bipartite graph of the miRNA-target association network. The schema of the heterogeneous network used in this paper is illustrated in Figure 1. The network contains two types of objects, miRNA and its targets. A meta-path $P$ is defined at the object type level and is denoted in the form of $A_1 \rightarrow A_2 \rightarrow \cdots A_{n-1} \rightarrow A_n$, where $A_i$ represent the object of type.

*3.2. Relatedness Measure.* The RM measure [36] is a path-constrained measure and it can calculate the relatedness of heterogeneous objects with the same or different types in a uniform framework. It has been proven that RM has some good properties, such as symmetric and self-maximum, and has shown its potential to mining valuable information in heterogeneous network. Therefore, here we use RM measure to calculate the relatedness between miRNA and its targets. RM measure is based on the Linkage Homophily Principle defined as follows.

*Linkage Homophily Principle.* Two nodes are more likely to be directly linked if most of their respective similar nodes are linked.

In general, the computing of nodes similarity is based on their neighbors. However, in heterogeneous networks, the same type similar nodes can be linked by heterogeneous nodes through composite paths. For example, two similar genes can be connected by a common miRNA, "gene → miRNA → gene." Thus, we can utilize meta-path to extract the generalized neighbor and define the similarity. Here, we first extract the meta-path that connects the source node and target node. We take source node $v_i^p$ and meta-path $P_{iis}$ as an example. The neighbors of node $v_i^p$ based on $P_{iis}$ are the nodes of type $i$ that linked to $v_i^p$ by $P_{iis}$, denoted as $N_i^p$. Similarly, we can get the generalized neighbors of target node $v_j^q$ and meta-path $P_{jjt}$, denoted as $N_j^q$. Then, we can use the connectivity between $N_i^p$ and $N_j^q$ to calculate the link's existence probability between nodes $v_i^p$ and $v_j^q$.

Suppose $\mathrm{RMP}_{iis}$ is the similarity matrix of $i$th type node along the meta-path $P_{iis}$. Similarity, $\mathrm{RMP}_{jjt}$ represents the similarity matrix of $j$th type node along the meta-path $P_{jjt}$. In general, similarity can be calculated by the path counts. Expected path number is the number where all of the links may exist from node of type $i$ to node of type $j$. Let meta-path $P_{ijst} = \{R_{k_0k_1}, R_{k_1k_2}, \ldots, R_{k_{n-1}k_n}\}$, $k_0 = i$, and $k_n = j$; then the expected path number $\mathrm{RMP}_{ijst}$ is computed as follows:

$$\mathrm{RMP}_{ijst} = \prod_{p=1}^{n} w_{k_{p-1}k_p} = \mathrm{RMP}_{iis} \times W_{ij} \times \mathrm{RMP}_{jjt}. \tag{1}$$

Here, $P_{ijst}$ is a meta-path composed of $P_{iis}$, $R_{ij}$, and $P_{jjt}$; $\mathrm{RMP}_{ijst}$ is a matrix whose size is $|V_i| \times |V_j|$. The computation of $\mathrm{RMP}_{iis}$ (or $\mathrm{RMP}_{jjt}$) is similar to the computation of $\mathrm{RMP}_{ijst}$.

Now the relatedness between nodes of type $i$ and nodes of type $j$ along the meta-path $P_{ijst}$ can be formulated as follows:

$$\begin{aligned} \mathrm{RM}_{ijst} &= \frac{\mathrm{RMP}_{ijst}}{\mathrm{RMP}_{iis} \times \mathbf{1} \times \mathrm{RMP}_{jjt}} \\ &= \frac{\mathrm{RMP}_{iis} \times W_{ij} \times \mathrm{RMP}_{jjt}}{\mathrm{RMP}_{iis} \times \mathbf{1} \times \mathrm{RMP}_{jjt}}. \end{aligned} \tag{2}$$

Here $\mathbf{1}$ is a matrix in which all the elements are 1 and the size of is $|V_i| \times |V_j|$. Similarly, $\mathrm{RM}_{ijst}$ is also a $|V_i| \times |V_j|$ matrix and $\mathrm{RM}_{ijst}^{pq}$ is the relatedness measured between $v_i^p$ and $v_j^q$ following $P_{ijst}$.

*3.3. Construction of the Feature Vector.* We can get the relatedness between miRNAs and their targets as described in Section 3.2. Now we get the feature vector as follows:

(1) Extract meta-path $P_{ii}$ of $i$th type node and $P_{jj}$ of $j$th type node.

(2) Compute the similarity based on any pair of meta-paths $P^{ii}$ and $P^{jj}$ and then get the feature vector.

In RMLM, the feature vector between miRNA $i$ and gene $j$ is defined as

$$\phi_{ij} = (f_1, f_2, \ldots, f_n), \tag{3}$$

where $f_1$ to $f_n$ represent the different similarities of different meta-paths with different semantic meaning.

*3.4. Parameter Estimation.* As different meta-path corresponds to different relation graphs, the final result may be improved by combining these different graphs through different weights. Here, logistic regression and maximum-likelihood estimation (MLE) method can be employed to estimate the weight.

In this paper, we regard the issue of relationship prediction as a two-class classification problem by using Bayesian analysis and logistic regression. Based on logistic regression and under general assumption [31, 32], the posterior probability of a specific relation can be formulated as follows:

$$p(x_i = 1 \mid \varphi_i, \omega) = \frac{\exp(\omega^T \varphi_i)}{\exp(\omega^T \varphi_i) + 1}, \tag{4}$$

$$p(x_i = 0 \mid \varphi_i, \omega) = \frac{1}{\exp(\omega^T \varphi_i) + 1}. \tag{5}$$

Here $\omega$ is a weight vector served as parameters and $\varphi_i$ is the feature vector of the link $x_i$. Then, MLE method can be employed to estimate the parameter vector $\omega$. The likelihood function can be written as

$$L(\omega; x_1, x_2, \ldots, x_N) = \prod_{i=1}^{N} p(x_i \mid \varphi_i, \omega). \tag{6}$$

Here $x_i$ is the link to calculate and $N$ is the number of links, $\varphi_i$ is the feature vector that is calculated according to RM, and $\omega$ is the weight vector of the feature according to different meta-path. The log likelihood of (6) is

$$\begin{aligned} &\ln L(\omega; x_1, x_2, \ldots, x_N) \\ &\quad = \sum_{i=1}^{N} \left[ x_i \omega^T \varphi_i - \ln\left(1 + \exp\left(\omega^T \varphi_i\right)\right)\right]. \end{aligned} \tag{7}$$

The log likelihood (7) is a convex function [40]. Hence, we can find a unique global optimal solution by solving a convex optimization problem.

*3.5. Final Score.* The logistic regression based algorithm returns a set of posterior probabilities. One can directly use those probabilities to make decision. However, the posterior probabilities do not always work well because it is difficult to set a threshold for a relation between miRNA and its target. Here, we utilize a percentage value as the final score to evaluate the strength of the relation between a miRNA and its target. The final score is calculated as follows:

$$q_i = \frac{\left|\{j \mid p_i \geq p_j\}\right|}{n}, \quad i = 1, 2, \ldots, n. \tag{8}$$

Here $\{p_1, p_2, \ldots, p_n\}$ is the posterior probabilities of any association, and $q_i$ is the top percentage value of $p_i$ among all those posterior probabilities. The larger the final score is, the more likely the association exists.

*3.6. Integration of Sequence Information.* In RMLMSe, we integrate sequence information to improve the performance of the RMLM. Here, we use sequence information from database TargetScan, miRanda, and PITA. As they have a relatively high false-positive rate, we only download conserved targets information and select the data whose Pct > 0.9 from TargetScan, mirSVR > 0.6 from miRanda, and data in PITATOP to improve the reliability of the regulation relationships. Sequence information from these databases acts as new features in feature vector used in RMLMSe.

Taking interaction between miRNA $i$ and gene $j$ as an example, its feature vector can be written as

$$\phi_{ij} = \left(f_1, f_2, \ldots, f_n, f_m, f_{m+1}, f_{m+2}\right). \quad (9)$$

Here $f_1$ to $f_n$ represent the different feature of different meta-paths and $f_m$, $f_{m+1}$, and $f_{m+2}$ represent the feature of sequence information from TargetScan, miRanda, and PITA, respectively.

*3.7. Algorithm.* The process description of RMLM and RMLMSe is given as follows.

*Input*. The disease set $d_i$ of each miRNA $i$ from HMDD and DAG $g_j$ of each disease $j$ from MeSH, the protein interaction matrix SP, and the miRNA-protein matrix MP.

*Output*. The vector of final score for each unknown interaction between miRNA and its targets.

(1) Calculate the miRNA-miRNA functional similarity matrix SM as described in Section 3.1.1.

(2) Extract meta-path $P_{ii}$ of $i$th type node and $P_{jj}$ of $j$th type node. We set the max length of meta-path between the same type node as (3).

(3) Concatenate $P_{iis}$ ($s$th meta-path of $P_{ii}$), $R_{ij}$, and $P_{jjt}$ ($t$th meta-path of $P_{jj}$) in sequence to compose a meta-path $P_{ijst}$ going from the source nodes of type $i$ to target nodes of type $j$. Then, the relatedness between miRNA and its target based on meta-path $P_{ijst}$ is calculated according to (2).

(4) Calculate the different similarity of different meta-path and get the feature vector of each interaction. The feature vectors used in RMLM and RMLMSe are described in Sections 3.3 and 3.5.

(5) Estimate parameters $\omega$ by maximizing the log likelihood $\ln L(\omega; x_1, x_2, \ldots, x_N)$ in (7) based on $x_i$ and $\varphi_i$, $x_i$ is the link to be calculated, and $N$ is the number of links.

(6) Calculate the probability for each unknown interaction according to (4) by using $\omega$ and feature vector.

(7) Calculate the final score according to (8).

## 4. Results

### 4.1. Datasets

*The Human MicroRNA Disease Database.* HMDD [41] provides a comprehensive resource of experimentally verified miRNA-disease associations. We can get the information through a website at http://www.cuilab.cn/hmdd. The database (in June 2014) contains 5100 associations between 491 miRNAs and 326 diseases. In this paper, we first analyse the global network. Then, we analyse another three diseases, Ovarian Neoplasms (OV), Lung Neoplasms (Lung), and Breast Neoplasms (Breast). The miRNAs associated with OV, Lung, and Breast are 114, 132, and 202, respectively.

*The Protein-Protein Interaction Database.* The PPI network was constructed by combining DNA-protein data from TRANSFAC [42] and protein interaction data obtained from Bossi and Lehner [43], respectively. The database contains 13306 proteins and 157426 interactions between proteins.

*Experimentally Validated miRNA-mRNA Interaction Databases*. The posttranscriptional regulatory knowledge is obtained from miRNA-target database miRTarbase *v6.1*. When mapping onto our miRNA-target matrix, it retains 111770 interactions. We can get the information through a website at (http://mirtarbase.mbc.nctu.edu.tw/).

*Predicted miRNA-mRNA Interaction Database.* We also utilize sequence information in database TargetScan *v7.0*, miRanda released at 2010, and PITA *v6*. These databases are available online at http://www.targetscan.org/, http://www.microrna.org/, and http://genie.weizmann.ac.il/pubs/mir07/, respectively.

*4.2. Comparisons with Other Methods.* To compare the performance of RMLM and RMLMSe, we applied RLSMDA [44] and RM [36] to the same testing data. RLSMDA was introduced to predict disease-miRNA association. We encoded RLSMDA in MATLAB according to the derivation process of the authors. Here, we set $\omega$ used in RLSMDA as 0.5. RM was implemented in MATLAB with source code available from authors personal homepage. RM is the measurement used to calculate the similarity of objects in heterogeneous networks. Here, the sum of the different similarities corresponding to different meta-paths is utilized to predict the miRNA-gene associations. All experiments are carried on a Windows 7 professional computer (Inter(R) Xeon(R) CPU, 2.93 GHz, 56 G RAM, 64-bit OS). The performance of each method is evaluated by fivefold cross validation. First, all known miRNA-target associations were split into five sets of the same size randomly: one set was set aside as the test set and the other four sets were used as train sets. The experiment was repeated five times so that each set was hidden once and each hidden miRNA-target pair obtained a predict relevance score. The ROC (receiver operating characteristic) curve was calculated according to the various TPR (true-positive rate) and the various FPR (false-positive rate) through a varying threshold. The area under the ROC curve (AUC) is employed to show the overall performance of methods. We can see from Figure 2 that RMLM and RMLMSe always work better than RLSMDA and RM. There is only slight improvement when sequence information is employed, where the AUC score increases from 0.8919 to 0.9033. This may have two reasons. First, the performance of the RMLM already achieves a very high AUC score and there is only a little room for it to be further improved by using additional prior information. Second, the amount of the sequence information mapped onto the miRNA-target matrix is little; for example, when TargatScan, miRanda, and PITA mapped onto the miRNA-target matrix, they leave 16,7403, 10,4631, and 13,7229 interactions, about 1.6~2.6% of the entire size of the miRNA-target matrix MP (a 491 × 13306 matrix). Although the improvement of the sequence

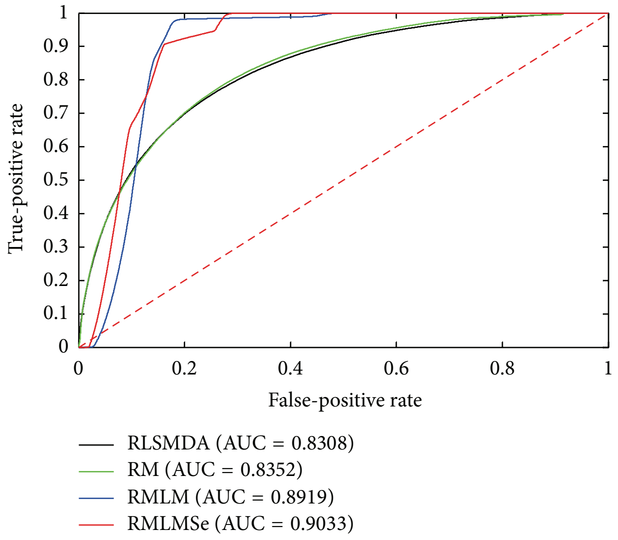

Figure 2: The ROC curve of the global network.

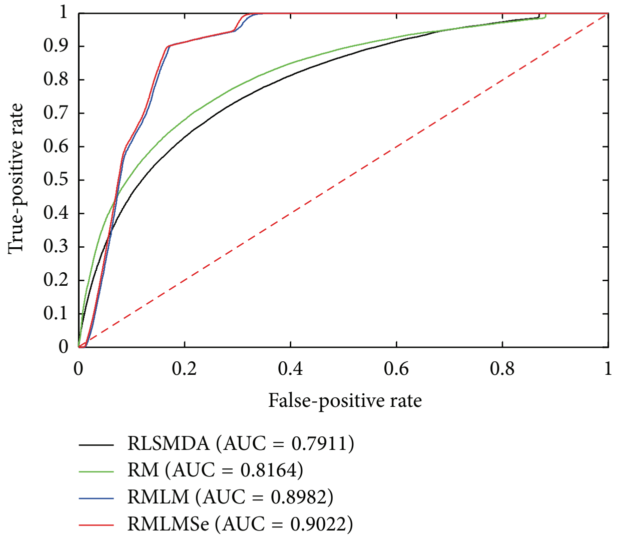

Figure 3: The ROC curve of the OV network.

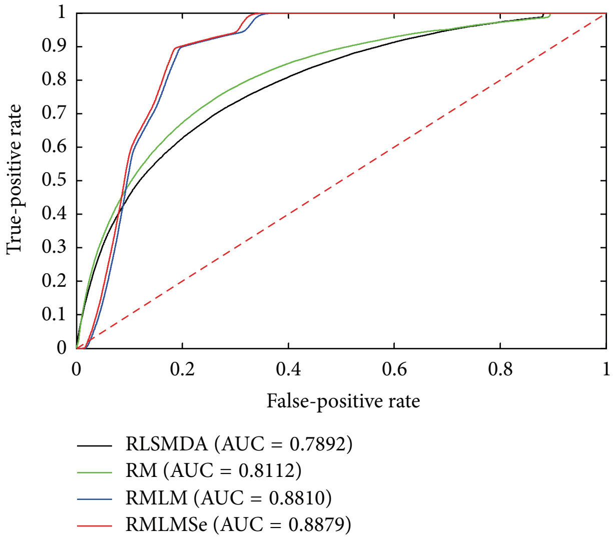

Figure 4: The ROC curve of the Lung network.

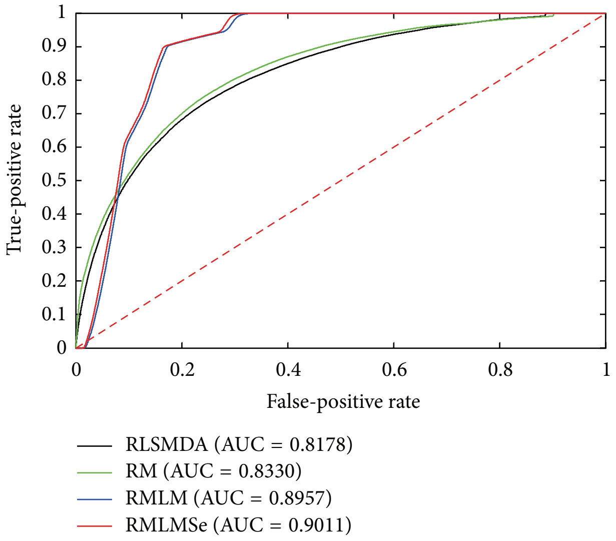

Figure 5: The ROC curve of the Breast network.

information is not significant, the increased AUC score still indicates that additional knowledge is helpful for improving the prediction performance as any prior knowledge, such as sequence information, Go Ontology annotations, gene copy numbers, and gene methylation, related to miRNA-target associations can be employed to predict associations. Figures 3, 4, and 5 are the result when we execute the methods on OV, Lung, and Breast database, respectively. The results are similar to Figure 2. RMLM and RMLMSe always work better than RLSMDA and RM, and RMLMSe only have a slight improvement than RMLM.

*4.3. The Number of Links Predicted by Our Methods.* Here, we present the number of interactions predicted based on different thresholds in RMLM and RMLMSe. As shown in Table 1, the numbers of interactions predicted in RMLM are higher than in RMLMSe among all of the threshold. This can further indicate the performance improvement in RMLMSe. In future, we can utilize the associations predicted by our method to construct miRNA-target regulatory network and extract regulatory modules and hub nodes.

*4.4. Functional Validation of mRNAs.* When we get the result of the global dataset, we compute every mRNA score and extract the top 250 mRNAs to carry on the pathway enrichment analysis with the focus on KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways (adjusted $p$ value < 0.05). In this paper, $p$ value calculated by hypergeometric test is a statistical value that represents the significant enrichment of pathways. The smaller the $p$ value is, the more significant the pathway enrichment is. As shown in Table 2, many of the KEGG pathways are highly related to many cancers and

Table 1: The number of links predicted by our methods based on different thresholds.

| Database | Methods | Validated | Th ≥ 0.9 | Th ≥ 0.8 | Th ≥ 0.7 | Th ≥ 0.6 | Th ≥ 0.5 |
|---|---|---|---|---|---|---|---|
| Global | RMLM | 11,1770 | 17,2912 | 20,4894 | 23,4327 | 26,5883 | 79,8049 |
|  | RMLMSe | 11,1770 | 17,6625 | 21,0909 | 24,2946 | 28,1782 | 80,7688 |
| OV | RMLM | 4,2730 | 5,3683 | 5,9580 | 6,4676 | 6,9759 | 23,3784 |
|  | RMLMSe | 4,2730 | 5,3891 | 5,9954 | 6,5526 | 7,1565 | 23,4562 |
| Lung | RMLM | 4,7764 | 5,8511 | 6,4339 | 6,9397 | 7,4816 | 24,5323 |
|  | RMLMSe | 4,7764 | 5,8870 | 6,4881 | 7,0437 | 7,9293 | 24,6261 |
| Breast | RMLM | 6,4403 | 8,6555 | 9,8883 | 10,9659 | 12,0730 | 36,4375 |
|  | RMLMSe | 6,4403 | 8,6690 | 9,9540 | 11,1719 | 12,6556 | 36,6573 |

The "validated" column is the number of links validated in database miRTarbase *v6.1* and "Th" represents the threshold.

Table 2: In RMLMSe, the enrichment KEGG pathways of global dataset.

|  | Enrichment KEGG pathways | $p$ value |
|---|---|---|
| 1 | p53 signaling pathway | $4.27E - 10$ |
| 2 | Chronic myeloid leukemia | $8.80E - 10$ |
| 3 | Bladder cancer | $3.24E - 09$ |
| 4 | Glioma | $6.03E - 09$ |
| 5 | Melanoma | $1.35E - 08$ |
| 6 | Pathways in cancer | $2.34E - 08$ |
| 7 | Prostate cancer | $1.01E - 07$ |
| 8 | Cell cycle | $1.61E - 07$ |
| 9 | Small cell lung cancer | $9.71E - 07$ |
| 10 | Pancreatic cancer | $3.26E - 06$ |

The $p$ values have been obtained through hypergeometric test.

Table 3: In RMLMSe, the enrichment KEGG pathways of lung dataset.

|  | Enrichment KEGG pathways | $p$ value |
|---|---|---|
| 1 | p53 signaling pathway | $5.15E - 10$ |
| 2 | Pathways in cancer | $3.11E - 08$ |
| 3 | Small cell lung cancer | $1.12E - 06$ |
| 4 | Non-small cell lung cancer | $1.04E - 05$ |
| 5 | Focal adhesion | $1.53E - 05$ |
| 6 | Neurotrophin signaling pathway | $1.81E - 04$ |
| 7 | Adherens junction | $6.05E - 04$ |
| 8 | ErbB signaling pathway | $1.34E - 03$ |
| 9 | Pathogenic *Escherichia coli* infection | $1.89E - 03$ |
| 10 | MAPK signaling pathway | $1.31E - 02$ |

The $p$ values have been obtained through hypergeometric test.

respective biological process, for instance, glioma, prostate cancer, and colorectal cancer. Furthermore, pathways in cancer are closely related to many cancers and P53 signaling pathways is proved to be related to the processes of cell division and DNA replication [45]. The result of Lung KEGG pathways is shown in Table 3. The pathway focal adhesion [46], adherens junction [47], and ErbB signaling pathway [48] are proved to be related to Lung.

## 5. Discussion and Conclusion

The rapid increase of various biological data provides challenges and opportunities for us to complete the global miRNA regulatory mechanism. In recent years, academics have made great efforts to predict miRNA targets. However, each method has its pros and cons, and the performance of a method varies on different datasets. Thus, how to get precise results is a long-time challenge for miRNA-target association prediction.

In this paper, two novel methods, RMLM and RMLMSe, were developed. In RMLM, we first construct miRNA-miRNA similarity matrix. Second, we use RM to evaluate the different relatedness between miRNAs and its target based on different meta-path and extract the feature vectors of links; different meta-path corresponds to different relation graphs; we can improve the performance by combining these different graphs through different weights of corresponding meta-paths. Third, logistic regression and MLE method were employed to estimate the weight. Here, the issue of relationship prediction is regarded as a two-class classification problem by using Bayesian analysis and logistic regression and then MLE method can be employed to estimate the parameter vector. Then, we estimate the posterior probabilities between miRNAs and its targets based on the feature vectors of links and the corresponding parameter vectors. Finally, the final scores are obtained by using the percentage values of individual posterior probabilities. In RMLMSe, we utilize more information such as sequence information from TargetSacn, miRanda, and PITA to improve the performance of the RMLM. The results showed that there are slight improvement when sequence information is integrated.

Compared with other methods, RMLM and RMLMSe proposed by us have higher AUC scores. Besides, we conduct pathway enrichment analysis and found many relevant pathways. These results indicate that our two methods were reasonable and credible.

The comparison results of RMLM and RMLMSe indicate that our methods have the capability to integrate more biological data, such as sequence data and gene copy number. Thus, with the rapid growth of the gene regulatory knowledge, our method can integrate more prior information to improve the prediction performance.

In addition, disease target inference [49, 50], disease-miRNA prioritization [51–54], and lncRNA-disease association prediction [55] are also the immediate areas of research focus to further study therapeutic strategy. Due to the scalability of the proposed methods, RMLM and RMLMSe could be applied to the different constructed heterogeneous networks to infer disease target, miRNA-disease association, and lncRNA-disease association, respectively. Moreover, the performance of our methods should be further evaluated after extending.

Of course, RMLM and RMLMSe also have some limitations that need to be improved in the future. Firstly, our methods utilize the network topology and known miRNA-gene associations to calculate the relatedness between miRNA and its target. It may cause bias to miRNA-gene pair which has more neighbor nodes. Furthermore, although the better performance is obtained by our methods on the whole, the predictive results should be further improved, especially for the small output. In the future, the prediction performance will be further improved by integrating more reliable biological data and obtaining more known miRNA-gene associations.

## Competing Interests

The authors declare that they have no competing interests.

## Acknowledgments

The research is supported by National Natural Science Foundation of China (Grant no. 61572180).

## References

[1] E. Wienholds and R. H. A. Plasterk, "MicroRNA function in animal development," *FEBS Letters*, vol. 579, no. 26, pp. 5911–5922, 2005.

[2] I. Alvarez-Garcia and E. A. Miska, "MicroRNA functions in animal development and human disease," *Development*, vol. 132, no. 21, pp. 4653–4662, 2005.

[3] W. C. S. Cho, "OncomiRs: the discovery and progress of microRNAs in cancers," *Molecular Cancer*, vol. 6, no. 1, article 60, pp. 1–7, 2007.

[4] F. Felicetti, M. C. Errico, L. Bottero et al., "The promyelocytic leukemia zinc finger-microRNA-221/-222 pathway controls melanoma progression through multiple oncogenic mechanisms," *Cancer Research*, vol. 68, no. 8, pp. 2745–2754, 2008.

[5] H. Guo, N. T. Ingolia, J. S. Weissman, and D. P. Bartel, "Mammalian microRNAs predominantly act to decrease target mRNA levels," *Nature*, vol. 466, no. 7308, pp. 835–840, 2010.

[6] N. Mercatelli, V. Coppola, D. Bonci et al., "The inhibition of the highly expressed mir-221 and mir-222 impairs the growth of prostate carcinoma xenografts in mice," *PLoS ONE*, vol. 3, no. 12, Article ID e4029, 2008.

[7] G. T. Huang, C. Athanassiou, and P. V. Benos, "MirConnX: condition-specific mRNA-microRNA network integrator," *Nucleic Acids Research*, vol. 39, no. 2, pp. W416–W423, 2011.

[8] B. Liu, J. Li, A. Tsykin, L. Liu, A. B. Gaur, and G. J. Goodall, "Exploring complex miRNA-mRNA interactions with Bayesian networks by splitting-averaging strategy," *BMC Bioinformatics*, vol. 10, article 408, 2009.

[9] S. R. A. Fisher, R. A. Fisher, S. Genetiker et al., *The Design of Experiments*, 1960.

[10] S.-D. Hsu, Y.-T. Tseng, S. Shrestha et al., "miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions," *Nucleic Acids Research*, vol. 42, no. 1, pp. D78–D85, 2014.

[11] R. F. Service, "Biology's dry future," *Science*, vol. 342, no. 6155, pp. 186–189, 2013.

[12] J. C. Huang, T. Babak, T. W. Corson et al., "Using expression profiling data to identify human microRNA targets," *Nature Methods*, vol. 4, no. 12, pp. 1045–1049, 2007.

[13] T. De Bie, L.-C. Tranchevent, L. M. M. van Oeffelen, and Y. Moreau, "Kernel-based data fusion for gene prioritization," *Bioinformatics*, vol. 23, no. 13, pp. i125–i132, 2007.

[14] B. P. Lewis, I.-H. Shih, M. W. Jones-Rhoades, D. P. Bartel, and C. B. Burge, "Prediction of mammalian microRNA targets," *Cell*, vol. 115, no. 7, pp. 787–798, 2003.

[15] B. P. Lewis, C. B. Burge, and D. P. Bartel, "Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets," *Cell*, vol. 120, no. 1, pp. 15–20, 2005.

[16] A. J. Enright, B. John, U. Gaul, T. Tuschl, C. Sander, and D. S. Marks, "MicroRNA targets in *Drosophila*," *Genome Biology*, vol. 5, no. 1, article R1, 2003.

[17] A. Krek, D. Grün, M. N. Poy et al., "Combinatorial microRNA target predictions," *Nature Genetics*, vol. 37, no. 5, pp. 495–500, 2005.

[18] M. Reczko, M. Maragkakis, P. Alexiou, I. Grosse, and A. G. Hatzigeorgiou, "Functional microRNA targets in protein coding sequences," *Bioinformatics*, vol. 28, no. 6, pp. 771–776, 2012.

[19] M. Kertesz, N. Iovino, U. Unnerstall, U. Gaul, and E. Segal, "The role of site accessibility in microRNA target recognition," *Nature Genetics*, vol. 39, no. 10, pp. 1278–1284, 2007.

[20] P. Sethupathy, M. Megraw, and A. G. Hatzigeorgiou, "A guide through present computational approaches for the identification of mammalian microRNA targets," *Nature Methods*, vol. 3, no. 11, pp. 881–886, 2006.

[21] H. Liu, A. R. Brannon, A. R. Reddy et al., "Identifying mRNA targets of microRNA dysregulated in cancer: with application to clear cell renal cell carcinoma," *BMC Systems Biology*, vol. 4, article 51, 2010.

[22] I. Van der Auwera, R. Limame, P. Van Dam, P. B. Vermeulen, L. Y. Dirix, and S. J. Van Laere, "Integrated miRNA and mRNA expression profiling of the inflammatory breast cancer subtype," *British Journal of Cancer*, vol. 103, no. 4, pp. 532–541, 2010.

[23] S. Kim, M. Choi, and K.-H. Cho, "Identifying the target mRNAs of microRNAs in colorectal cancer," *Computational Biology and Chemistry*, vol. 33, no. 1, pp. 94–99, 2009.

[24] Y. Lu, Y. Zhou, W. Qu, M. Deng, and C. Zhang, "A Lasso regression model for the construction of microRNA-target regulatory networks," *Bioinformatics*, vol. 27, no. 17, pp. 2406–2413, 2011.

[25] A. Muniategui, R. Nogales-Cadenas, M. Vázquez et al., "Quantification of miRNA-mRNA interactions," *PLoS ONE*, vol. 7, no. 2, Article ID e30766, 2012.

[26] N. Su, Y. Wang, M. Qian, and M. Deng, "Predicting MicroRNA targets by integrating sequence and expression data in cancer," in *Proceedings of the 5th IEEE International Conference on Systems Biology (ISB '11)*, pp. 219–224, Zhuhai, China, September 2011.

[27] T. D. Le, L. Liu, A. Tsykin et al., "Inferring microRNA-mRNA causal regulatory relationships from expression data," *Bioinformatics*, vol. 29, no. 6, pp. 765–771, 2013.

[28] F. C. Stingo, Y. A. Chen, M. Vannucci, M. Barrier, and P. E. Mirkes, "A Bayesian graphical modeling approach to microRNA regulatory network inference," *The Annals of Applied Statistics*, vol. 4, no. 4, pp. 2024–2048, 2010.

[29] X. Chen, F. J. Slack, and H. Zhao, "Joint analysis of expression profiles from multiple cancers improves the identification of microRNA-gene interactions," *Bioinformatics*, vol. 29, no. 17, pp. 2137–2145, 2013.

[30] A. Muniategui, J. Pey, F. J. Planes, and A. Rubio, "Joint analysis of miRNA andmRNA expression data," *Briefings in Bioinformatics*, vol. 14, no. 3, Article ID bbs028, pp. 263–278, 2013.

[31] F. Tai and W. Pan, "Incorporating prior knowledge of predictors into penalized classifiers with multiple penalty terms," *Bioinformatics*, vol. 23, no. 14, pp. 1775–1782, 2007.

[32] Z. Tian, T. Hwang, and R. Kuang, "A hypergraph-based learning algorithm for classifying gene expression and arrayCGH data with prior knowledge," *Bioinformatics*, vol. 25, no. 21, pp. 2831–2838, 2009.

[33] Z. Zhao, J. Wang, H. Liu, J. Ye, and Y. Chang, "Identifying biologically relevant genes via multiple heterogeneous data sources," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '08)*, pp. 839–847, August 2008.

[34] A. Kozomara and S. Griffiths-Jones, "miRBase: integrating microRNA annotation and deep-sequencing data," *Nucleic Acids Research*, vol. 39, supplement 1, pp. D152–D157, 2010.

[35] Y. Sun, B. Norick, J. Han, X. Yan, P. S. Yu, and X. Yu, "Pathselclus: integrating meta-path selection with userguided object clustering in heterogeneous information networks," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 7, no. 3, p. 11, 2013.

[36] B. Cao, X. Kong, and P. S. Yu, "Collective prediction of multiple types of links in heterogeneous information networks," in *Proceedings of the 14th IEEE International Conference on Data Mining (ICDM '14)*, pp. 50–59, Shenzhen, China, December 2014.

[37] Y. Sun, Y. Yu, and J. Han, "Ranking-based clustering of heterogeneous information networks with star network schema," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '09)*, pp. 797–806, Paris, France, July 2009.

[38] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu, "Pathsim: meta path-based top-k similarity search in heterogeneous information networks," *Proceedings of the VLDB Endowment*, vol. 4, no. 11, pp. 992–1003, 2011.

[39] D. Wang, J. Wang, M. Lu, F. Song, and Q. Cui, "Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases," *Bioinformatics*, vol. 26, no. 13, Article ID btq241, pp. 1644–1650, 2010.

[40] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, UK, 2004.

[41] Y. Li, C. Qiu, J. Tu et al., "HMDD v2.0: a database for experimentally supported human microRNA and disease associations," *Nucleic Acids Research*, vol. 42, no. 1, pp. D1070–D1074, 2014.

[42] V. Matys, O. V. Kel-Margoulis, E. Fricke et al., "TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes," *Nucleic Acids Research*, vol. 34, supplement 1, pp. D108–D110, 2006.

[43] A. Bossi and B. Lehner, "Tissue specificity and the human protein interaction network," *Molecular Systems Biology*, vol. 5, article 260, 2009.

[44] X. Chen and G.-Y. Yan, "Semi-supervised learning for potential human microRNA-disease associations inference," *Scientific Reports*, vol. 4, article 5501, 2014.

[45] S. L. Harris and A. J. Levine, "The p53 pathway: positive and negative feedback loops," *Oncogene*, vol. 24, no. 17, pp. 2899–2908, 2005.

[46] G. W. McLean, N. O. Carragher, E. Avizienyte, J. Evans, V. G. Brunton, and M. C. Frame, "The role of focal-adhesion kinase in cancer—a new therapeutic opportunity," *Nature Reviews Cancer*, vol. 5, no. 7, pp. 505–515, 2005.

[47] Q.-Y. Chen, D.-M. Jiao, L.-F. Wang et al., "Curcumin inhibits proliferation-migration of NSCLC by steering crosstalk between a Wnt signaling pathway and an adherens junction via EGR-1," *Molecular BioSystems*, vol. 11, no. 3, pp. 859–868, 2015.

[48] T. Yu, J. Li, M. Yan et al., "MicroRNA-193a-3p and -5p suppress the metastasis of human non-small-cell lung cancer by downregulating the ERBB4/PIK3R3/mTOR/S6K2 signaling pathway," *Oncogene*, vol. 34, no. 4, pp. 413–423, 2015.

[49] U. M. Singh-Blom, N. Natarajan, A. Tewari, J. O. Woods, I. S. Dhillon, and E. M. Marcotte, "Prediction and validation of gene-disease associations using methods inspired by social network analyses," *PLoS ONE*, vol. 8, no. 5, Article ID e58977, 2013.

[50] X. Zeng, Y. Liao, Y. Liu, and Q. Zou, "Prediction and validation of disease genes using HeteSim Scores," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2016.

[51] X. Chen, M.-X. Liu, and G.-Y. Yan, "RWRMDA: predicting novel human microRNA-disease associations," *Molecular BioSystems*, vol. 8, no. 10, pp. 2792–2798, 2012.

[52] P. Xuan, K. Han, M. Guo et al., "Prediction of microRNAs associated with human diseases based on weighted k most similar neighbors," *PLoS ONE*, vol. 8, no. 8, Article ID e70204, 2013.

[53] X. Chen, C. Clarence Yan, X. Zhang et al., "RBMMMDA: predicting multiple types of disease-microRNA associations," *Scientific Reports*, vol. 5, Article ID 13877, 2015.

[54] X. Chen, C. C. Yan, X. Zhang et al., "WBSMDA: within and between Score for MiRNA-disease association prediction," *Scientific Reports*, vol. 6, Article ID 21106, 2016.

[55] X. Chen, C. C. Yan, X. Zhang, and Z.-H. You, "Long non-coding RNAs and complex diseases: from experimental results to computational models," *Briefings in Bioinformatics*, 2016.