lncRNA-Disease Association Prediction Based On Weight Matrix And Projection Score

Bo Wang (drbowang@163.com)
Qiqihar University
Chao Zhang
Qiqihar University
Xiao-xin Du
Qiqihar University
Jian-fei Zhang
Qiqihar University

Research Article

Keywords: lncRNA-miRNA association, miRNA-disease association, disease semantic similarity, Integrated lncRNA similarity, integrated disease similarity, Weight allocation algorithm, Projection score.

Posted Date: April 23rd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-428221/v1

lncRNA-Disease Association Prediction Based On Weight Matrix And Projection Score

Bo Wang[1*], Chao Zhang[1], Xiao-xin Du[1], and Jian-fei Zhang[1]

*Correspondence:
drbowang@163.com
1 College of computer and control engineering, Qiqihar University, Qiqihar 161006, People's Republic of China
Full list of author information is available at the end of the article

Abstract

Background: with the development of medical science, lncRNA, originally considered as a noise gene, has been found to participate in a variety of biological activities. Nowadays, more and more studies show that lncRNA is involved in various human diseases, such as gastric cancer, prostate cancer, lung cancer, etc. However, obtaining lncRNA-disease association only through biological experiments not only costs manpower and material resources, but also gains little. Therefore, it is very important to develop effective computational models for predicting lncRNA-disease association.

Results: In this paper, a new lncRNA-disease association prediction model LDAP-WMPS based on weight distribution and projection score is proposed. Based on the existing research results of disease semantic similarity, the integrated lncRNA similarity matrix and the integrated disease similarity matrix are calculated according to the disease semantic similarity and the association information between data. On this basis, the weight algorithm is combined with the improved projection algorithm to predict the lncRNA-disease association through the known lncRNA-miRNA association and miRNA-disease association. The simulation results show that under the loocv framework, the AUC of LDAP-WMPS can reach 0.8822. Better than the latest results. Through the case study of adenocarcinoma and colorectal cancer, it is proved that LDAP-WMPS can effectively infer lncRNA-disease association.

Conclusions: The simulation results show that LDAP-WMPS has good prediction performance, which is an important supplement to the research of lncRNA-disease association prediction without lncRNA-disease association data.

Keywords: lncRNA-miRNA association, miRNA-disease association, disease semantic similarity, Integrated lncRNA similarity, integrated disease similarity, Weight allocation algorithm, Projection score.

Background

According to the traditional central principle, RNA is divided into messenger RNA (mRNA) and non-coding RNA (ncRNA). Messenger RNA is the medium for DNA to be transcribed into protein, while non coding RNA has always been regarded as noise and has no real effect. However, the sequencing results show that in the whole human gene pool, less than 5% of DNA and RNA are involved in protein transcription, and other genes are involved in RNA transcription that cannot be encoded, that is, the number of non-coding RNA is far greater than that of coding RNA [1]. In 1998, two American scientists Andrew Farr and Craig Mello jointly published a paper on the discovery of RNA interference mechanism in the journal Nature. They believed that RNA interference

exists in all organisms, and RNA plays a regulatory role in gene expression [2], virus infection [3,4], immune system [5], etc., thus, bringing biological research into a new stage. After that, the research on ncRNA gradually increased, among which the research on long non coding RNA (lncRNA) is one of the hot topics. Long non coding RNA is a kind of non coding RNA whose nucleotide length is more than 200. In previous studies, it was considered to be the noise generated in the process of transcription [6, 7] Nowadays, lncRNA has been found to be involved in all aspects of cell life cycle, including transcription [8], cell differentiation [9], cell transport [10], apoptosis [11], metabolic process [12] and so on. Moreover, lncRNA has also been found to be associated with various human diseases [13], including leukemia [14,15], diabetes [16,17], prostate cancer [18,19], lung cancer [20,21], colon cancer [22,23], cardiovascular disease [24, 25] and so on. lncRNA participates in diseases through abnormal sequence and spatial structure, abnormal expression level and abnormal interaction with binding proteins, thus affecting human health [26,27]. Therefore, linking lncRNA with diseases can realize the early detection of diseases, the targeted treatment of diseases, and the systematic understanding of the etiological characteristics of complex diseases. Because of the complex relationship between lncRNA and diseases, it costs a lot of money and time to carry out the biological experiments related to lncRNA. Computer aided experiment has become an effective research method. Computer aided experiments can effectively predict the association between lncRNA and complex diseases. For the prediction results, the data sets in the open lncRNA database are used to verify. The prediction of lncRNA disease association is of great significance in biology, medicine and other fields. In the field of biology, computer-aided experiments can reduce the cost of experiments and improve the success rate of experiments; in the field of medicine, computer-aided experiments can help researchers identify lncRNAs related to various diseases and understand the pathogenesis of diseases at the molecular level, so as to effectively prevent and treat diseases.So far, the prediction models put forward by various experts and scholars can be divided into two categories. The first model relies only on miRNA-disease association information or lncRNA-disease association information. Specifically, we can predict the association between miRNA-diseases by the association information between miRNA-diseases, and predict the association between lncRNA-diseases through the association information between lncRNA-diseases. For example, in this study, Guang et al. Proposed a label propagation model with linear neighborhood similarity, called LPLNS, to predict the potential association between lncRNA and disease[28]. Based on disease semantic similarity and lncRNA-disease association information, Guang et al. Developed an NCPLDA model to predict the potential association between lncRNA and diseases through network consistency [29]. Gu et al. Proposed a method to infer the pairwise functional similarity and functional network of human miRNA based on the RNA of disease relationship structure, so as to infer the new potential function of miRNA or related diseases [30]. The other model is to integrate multiple data, collect multiple biological data such as lncRNA, miRNA, protein, disease and so on, and integrate these data into matrix or heterogeneous network to infer the potential relationship between lncRNA and disease. For example, Yu and Wang et al. Developed a NBCLDA model, which integrates a variety of organisms to construct a new tripartite network, including miRNA-disease, miRNA-lncRNA and lncRNA-disease association and interaction. Then, a quadruple network is constructed and naive Bayesian classifier is applied to predict [31]. Chen et al. Proposed a new prediction model called LRLSLDA by fusing the known phenome - lncRNAome network, disease similarity network and lncRNA similarity network by using Laplace regularized least squares [32].Yu et al. Proposed a novel model CFNBC, which combined collaborative filtering with naive Bayes, and used to infer the potential lncRNA-disease association by calculating the association score between lncRNA and disease [33].Lu et al. Developed a computational model called SIMCLDA using inductive matrix. The principle is to complete the disease interaction of missing lncRNA based on known interactions, lncRNA similarity data and disease similarity data [34].

However, most of the prediction of lncRNA-disease correlation needs to know the correlation between lncRNA-diseases. But the known association between lncRNA-diseases is quite rare. To solve the above problems, this paper proposes a lncRNA-disease association prediction model LDAP-WMPS based on weight matrix and projection score. The model uses the relatively perfect lncRNA-miRNA association data and miRNA-disease association data to predict lncRNA-disease association. The integrated lncRNA

similarity matrix and integrated disease similarity matrix were established by fusing various methods to calculate the similarity between lncRNA and disease. On this basis, the weight algorithm is improved and applied to the lncRNA-miRNA-disease triple network. Based on the network, a new lncRNA-disease weight matrix calculation method is proposed. Combined with the improved projection algorithm, the lncRNA-miRNA Association and miRNA-disease association are used to predict the lncRNA-disease association. The simulation results show that under the loocv framework, the AUC of LDAP-WMPS can reach 0.8822. Better than the latest. Taking adenocarcinoma and colorectal cancer as examples, it is proved that LDAP-WMPS can effectively infer the relationship between lncRNA and disease.

Results

Performance evaluation

We evaluated the performance of LDAP-WMPS model by using Leave-One-Out Cross Validation (LOOCV), and compared the results with other prediction models using LOOCV, and compare the results with other prediction models for LOOCV. In the LOOCV experiment, for each disease j in the disease data set, we successively remove a lncRNAs that is known to be associated with the disease j. This one is the test set. The correlation score calculated in the test set is compared with the given threshold, we can get true positive (TP), true positive (TP), true negative (TN) and false negative (FN) by calculating one by one. In order to obtain the receiver operating characteristic curve (ROC) and the area under the ROC curve (AUC) for intuitive evaluation. True positive rate (TPR) and false positive rate (FPR) were calculated:

$$TPR = \frac{TP}{TP + FN} \quad (1)$$

$$FPR = \frac{FP}{FP + TN} \quad (2)$$

Receiver operating characteristic curve (ROC) was drawn with True positive rate (TPR) and False positive rate (FPR), and area under ROC curve (AUC) was calculated.

Comparison with other advanced models

In order to prove the effectiveness of LDAP-WMPS model, we compare the LDAP-WMPS model with other three advanced models. The ROC curve and AUC area are obtained by applying four different models to the same dataset. After comparison, LDAP-WMPS model is slightly better than other methods in ROC curve, and AUC reaches 0.8822. The highest AUC of CFNBC [33], and NBCLDA [31] models were 0.8576 and 0.8521, respectively. The results show that our method is slightly better than that used in CFNBC. The results are shown in Table 1, figure 1 and figure 2.

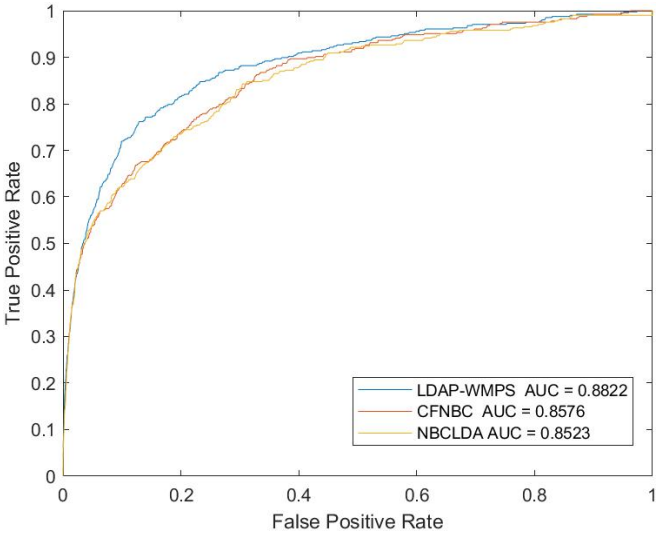

Figure1.The performance of LDAP-WMPS and others models in terms of ROC curves and AUCs based on 407 known lncRNA-disease associations under the framework of LOOCV

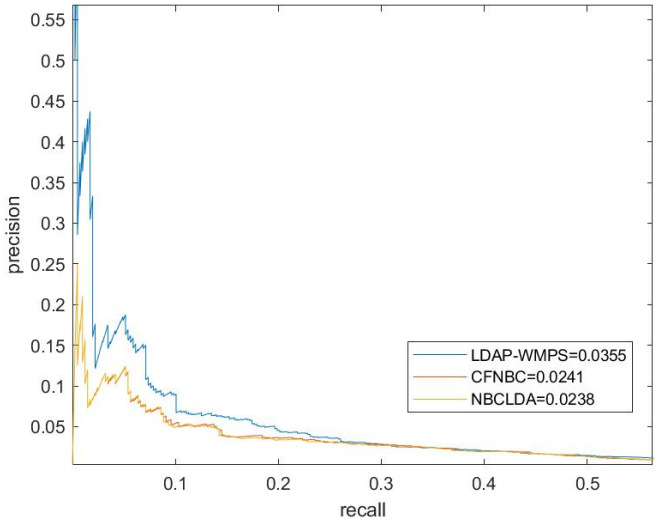

Figure2.The performance of LDAP-WMPS and others models in terms of PR curves and AUPRs based on 407 known lncRNA-disease associations under the framework of LOOCV

Table 1. AUC values of LDAP-WMPS model and other models under LOOCV framework under the same dataset

Method	AUC	AUPR
LDAP-WMPS	0.8822	0.0355
CFNBC	0.8576	0.0241
NBCLDA	0.8521	0.0238

Analysis of parameters

In this model, we introduce parameter ∂. The range of parameter ∂ are [0,1].When ∂ =0, only disease projection score is used for final score calculation; when ∂ =1, only lncRNA projection score is used for final score calculation The results are shown in figure 3 and figure 4.Obviously, when ∂ = 0.52, AUC reaches the highest value of 0.8822. In order to further prove the effectiveness of our lncRNA-disease weight matrix, we evaluated the model using weight matrix and the model not using weight matrix respectively, and the results are shown in Figure 5. It is obvious that our weight matrix effectively improves the prediction ability of the model.

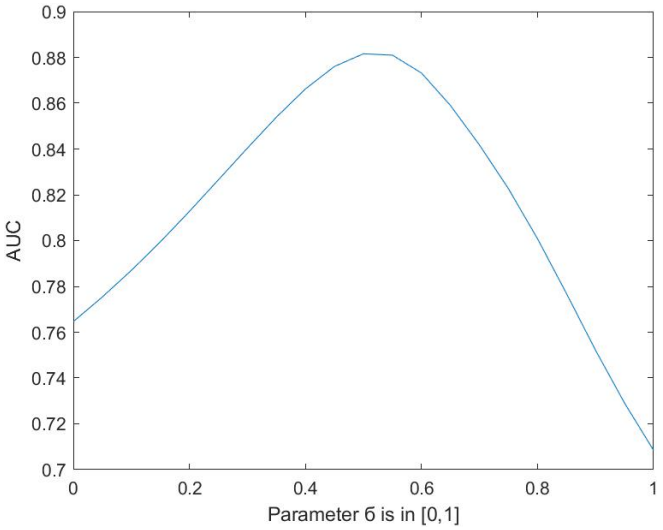

Figure 3. Transformation curve of parameter in the range of [0,1]

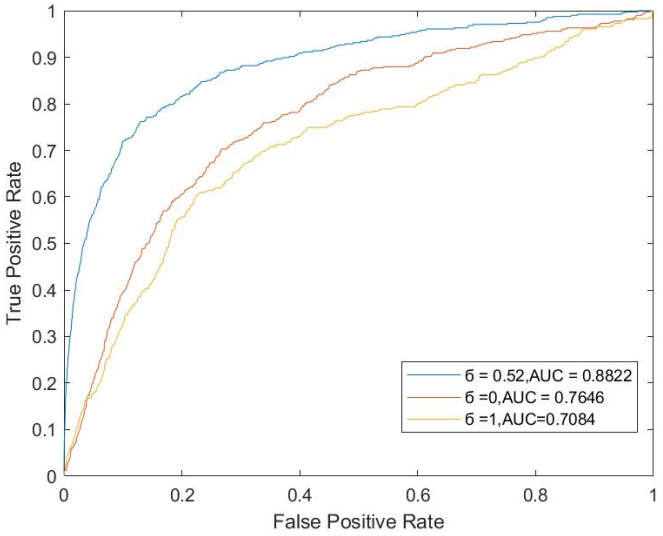

Figure 4. ROC calculated by fusion of lncRNA projection fraction and disease projection fraction was compared with ROC calculated by lncRNA projection fraction only and disease projection fraction only.

Case studies

Tumor refers to a new organism formed by the proliferation of local tissue cells under the action of various oncogenic factors, because this new organism is mostly space occupying massive protuberances, also known as vegetations. According to the cellular characteristics of tumors and the degree of harm to the body, tumors are divided into benign tumors and malignant tumors: benign tumors can be removed by surgery, and will not metastasize and relapse; malignant tumors, as we often call cancer, are easy to metastasize, difficult to cure by surgery, and there is still the possibility of recurrence after cure [35].In order to further prove the practicability of LDAP-WMPS in lncRNA-disease association prediction, we studied adenocarcinoma and colorectal cancer. The first 20 pieces of information about LDAP-WMPS predicting adenocarcinoma and colorectal cancer are shown in Table 2 and Table 3.

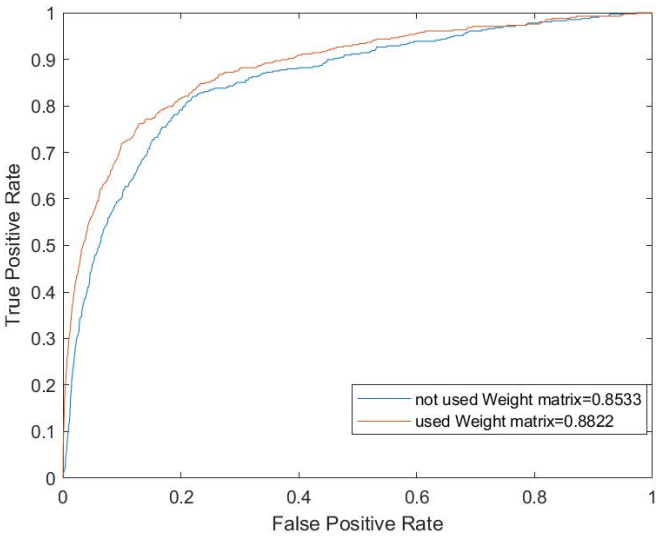

Figure 5. Comparison of ROC curve calculated with weight matrix and ROC curve calculated without weight matrix.

Table 2. Top 20 lncRNA of Colorectal Neoplasms predicted by LDAP-WMPS

lncRNA	Evidence (PMID)	Rank
XIST	28837144	1
MALAT1	25031737;21503572	3
DCP1A	29964337	4
KCNQ1OT1	16965397;11340379	5
NEAT1	30185232	8
OIP5-AS1	29773344	9
HCG18	31854468	10
FGD5-AS1	31332696	13
TUG1	31528224	14
RP4-773N10.5	31966592	15
SNHG16	32859986	18
GAS5	31619268	20

Colorectal cancer is a common cancer type. Its incidence rate and mortality rate are high in the world. In 2018 alone, the number of new cases reached nearly 2 million, and the number of deaths was nearly 900 thousand. Some data show that in the United States, about 5.2% of men and 4.8% of women are at risk of colorectal cancer, and the mortality caused by colorectal cancer is close to 33% [36]. Many studies have shown that lncRNA is closely related to colorectal cancer. In our prediction results, 12 of the first 20 lncRNAs associated with colorectal cancer have been proved by relevant medicine: lncRNA XIST expedites metastasis and modulates epithelial-mesenchymal transition in colorectal cancer[37];lncRNA SNHG16 promotes colorectal cancer cell proliferation, migration, and epithelial-mesenchymal transition through miR-124-3p/MCP-1[38];lncRNA MALAT1 promotes the colorectal cancer malignancy by increasing DCP1A expression and miR203 downregulation[39];The lncRNA HCG18 promotes the growth and invasion of colorectal cancer cells through sponging miR-1271 and upregulating MTDH[40];lncRNA FGD5-AS1 promotes colorectal cancer cell proliferation, migration, and invasion through upregulating CDCA7 via sponging miR-302e [41];Long non-coding RNA TUG1 mediates 5-fluorouracil resistance by acting as a ceRNA of miR-197-3p in colorectal cancer[42].

Table 3. Top 20 lncRNA of Adenocarcinoma predicted by LDAP-WMPS

lncRNA	Evidence (PMID)	Rank
XIST	28961027	1
MALAT1	31480991	3
DCP1A	25089265	4
RP6-24A23.7	28299977	5
KCNQ1OT1	30932685	7
HCG18	32559619	8
NEAT1	30036873	9
OIP5-AS1	32669972	10
CTB-89H12.4	26975529	12
FGD5-AS1	33416094	13
SNHG16	31580045	16
SENP3-EIF4A1	32602848	17
TUG1	29960845	18
LINC00662	33108738	20

Adenocarcinoma is a kind of lung cancer. It is the least related to smoking, accounting for 40% of primary Adenocarcinoma. Often located in the peripheral part of the lung, but also involving the pleura and the formation of associated scar ring and pleural effusion. Because of the invasive growth of adenocarcinoma, extensive resection should be performed. The rate of lymph node metastasis of adenocarcinoma is high, which can be as high as 36% - 47%. It is easy to relapse and has poor prognosis. Lin Guoji reported 68 cases of adenocarcinoma. The 5-year and 10-year cure rates were 43.9% and 29.0% respectively[43].In our prediction results, 14 of the first 20 lncRNAs associated with Adenocarcinoma have been proved by relevant medicine:lncRNA XIST promotes human lung adenocarcinoma cells to cisplatin resistance via let-7i/BAG-1 axis[44].lncRNA MALAT1 promotes gastric adenocarcinoma through the miR-181a-5p/AKT3 axis[45].lncRNA CTB-89H12.4 regulation of PTEN expression in prostate cancer[46].lncRNA HCG18 acted an oncogene in lung adenocarcinoma and enhanced lung adenocarcinoma progression by targeting miR-34a-5p/HMMR axis[47].lncRNA SNHG16 promotes cell proliferation and invasion in lung adenocarcinoma via sponging let-7a-5p[48].

Discussion

To explore the relationship between lncRNA and diseases is not only of great significance to the treatment of diseases, but also helpful to explore the mystery of human body. Using artificial intelligence to mine the existing medical data can not only improve the utilization rate of data, but also speed up the process of medical intelligence. In this study, we propose a computational model LDAP-WMPS. In this model, we propose a weight allocation algorithm based on lncRNA-miRNA-disease triple network, and on this basis, we propose a lncRNA disease association weight calculation method, and combine the lncRNA disease weight matrix with the improved projection algorithm to calculate the relationship between each lncRNA and disease the interaction between lncRNA and disease information can be obtained. Compared with the other three models, LDAP-WMPS is slightly better in AUC. 12 of the first 20 lncRNAs have been confirmed to predict the relationship between adenocarcinoma and colorectal cancer, which also proves the reliability of LDAP-WMPS. In addition, our model is based on the lncRNA and miRNA Association and miRNA-disease association to achieve the prediction of lncRNA-disease association. Through

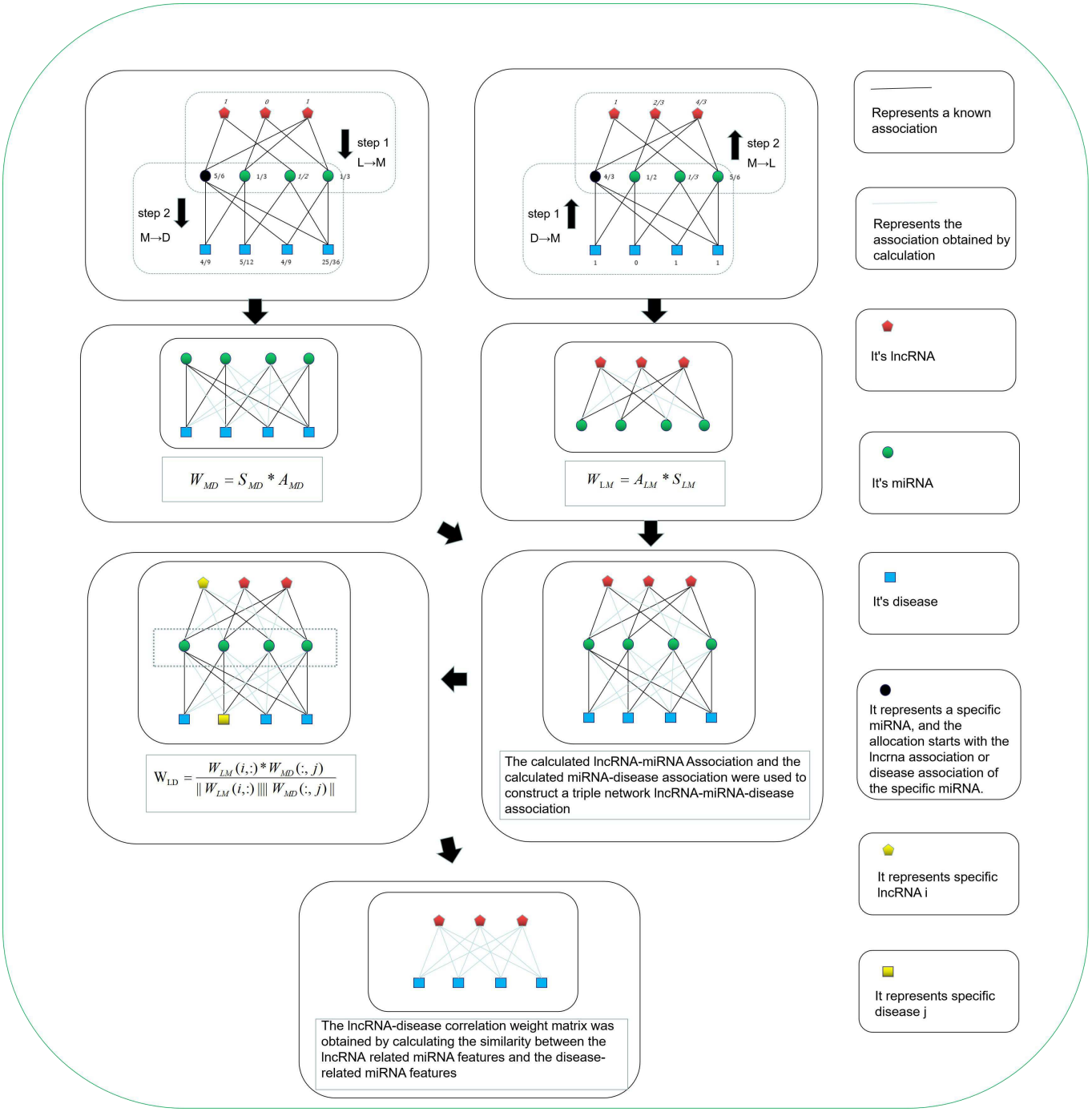

Figure 6. Flow chart of lnRNA-disease association weight matrix construction

the current relatively perfect lncRNA-miRNA association data set and miRNA-disease association data set to predict lncRNA-disease association can effectively avoid the current lncRNA-disease association based on lncRNA-disease association data Lack of lncRNA-disease association data in data prediction.

Conclusion

In this manuscript, we propose a new computing model LDAP-WMPS. Our main contributions are as follows: 1. We propose an integrated lncRNA similarity calculation method and an integrated disease similarity calculation method. 2. We propose a weight assignment algorithm for lncRNA-miRNA-disease triple network. 3. Based on the weight distribution of lncRNA-miRNA-disease triple network, a method of lncRNA-disease weight calculation is proposed. 4. Improve the existing consistency projection scoring formula. 5. lncRNA-disease association can be predicted by LDAP-WMPS without relying on the known lncRNA disease association data.

Method

lncRNA- disease association data set、miRNA- disease association data set、lncRNA-miRNA Association data Set

The known lncRNA-disease association dataset is download from mndrv2.0 database (2017 Edition) [49]. The known miRNA-disease association datasets are download from HMDD database (2018 Edition) [50]. The known lncRNA-miRNA Association dataset is download from Starbase v2.0 database (2015 Edition) [51]. After data cleaning and name unification, we get three datasets D_{LM} , D_{MD} , D_{LD} .There are 1089 different kinds of lncRNA and 246 different kinds of miRNA in D_{LM} data set,246 different types of miRNA and 373 different types of diseases in D_{MD} data set, and 1089 different types of lncRNA and 373 different types of diseases in D_{LD} data set. where D_{LD} data set is not used as training set, but only as test set the D_{MD} and D_{LM} data sets are analyzed and transformed into adjacency matrix. Taking lncRNA-miRNA association data set as an example, the adjacency matrix A_{LM} is constructed. The type of lncRNA was listed as the row, and the number of miRNA species was listed as the column. If the miRNA in row j interacts with lncRNA in column i, then $A_{LM}(i,j)=1$. On the contrary, $A_{LM}(i,j)=0$. Similarly, adjacency matrices A_{MD} can be constructed.

Cosine similarity for diseases

The cosine similarity for disease between miRNA disease adjacency matrix was calculated：

$$CD(i,j)=\frac{A_{MD}(:,i)*A_{MD}(:,j)}{\|A_{MD}(:,i)\|\|A_{MD}(:,j)\|} \tag{3}$$

Where $A_{MD}(:,i)$ is the i-th column vector in the adjacency matrix of miRNA and disease, which represents the association feature of disease i.

Jaccard similarity for diseases

The calculation of similarity is an important part of gene association prediction. At present, the methods of similarity calculation in most articles include Gauss interactive calculation of similarity. Compared with the past, we use Jaccard similarity to calculate. The Jaccard similarity for disease between miRNA disease adjacency matrix was calculated：

$$JD(i,j)=\frac{A_{MD}(:,i)\cap A_{MD}(:,j)}{A_{MD}(:,i)\cup A_{MD}(:,j)} \tag{4}$$

Where $A_{MD}(:,i)$ is the i-th column vector in the adjacency matrix of miRNA and disease, which represents the association feature of disease I. Similarly, $A_{MD}(:,i)$ represents the association feature of disease i; $A_{MD}(:,i)\cap A_{MD}(:,j)$ is the number of miRNAs associated with disease i and disease j, $A_{MD}(:,i)\cup A_{MD}(:,j)$ is the sum of miRNAs related to disease i and disease j

Integrated disease semantic similarity matrix

Integrated disease semantic similarity DS and cosine similarity CD for diseases:

$$IDS(i,j)=\begin{cases} CD(i,j) & if\ CD(i,j)\neq 0; \\ JD(i,j) & if\ CD(i,j)=0; \end{cases} \tag{5}$$

Cosine similarity for lncRNA

The cosine similarity for lncRNA between lncRNA-miRNA adjacency matrix was calculated：

$$CL(i,j)=\frac{A_{LM}(i,:)*A_{LM}(j,:)}{\|A_{LM}(i,:)\|\|A_{LM}(j,:)\|} \tag{6}$$

Where $A_{LM}(i,:)$ is the i-th column vector in the adjacency matrix of lncRNA and miRNA, which represents the association feature of lncRNA i.

Jaccard similarity for lncRNA

The Jaccard similarity for lncRNA between lncRNA-miRNA adjacency matrix was calculated：

$$JL(i,j)=\frac{A_{LM}(i,:)\cap A_{LM}(j,:)}{A_{LM}(i,:)\cup A_{LM}(j,:)} \tag{7}$$

Where $A_{LM}(i,:)$ is the i-th column vector in the adjacency matrix of lncRNA and miRNA, which represents the association feature of lncRNA I. Similarly, $A_{LM}(j,:)$ represents the association feature of lncRNA

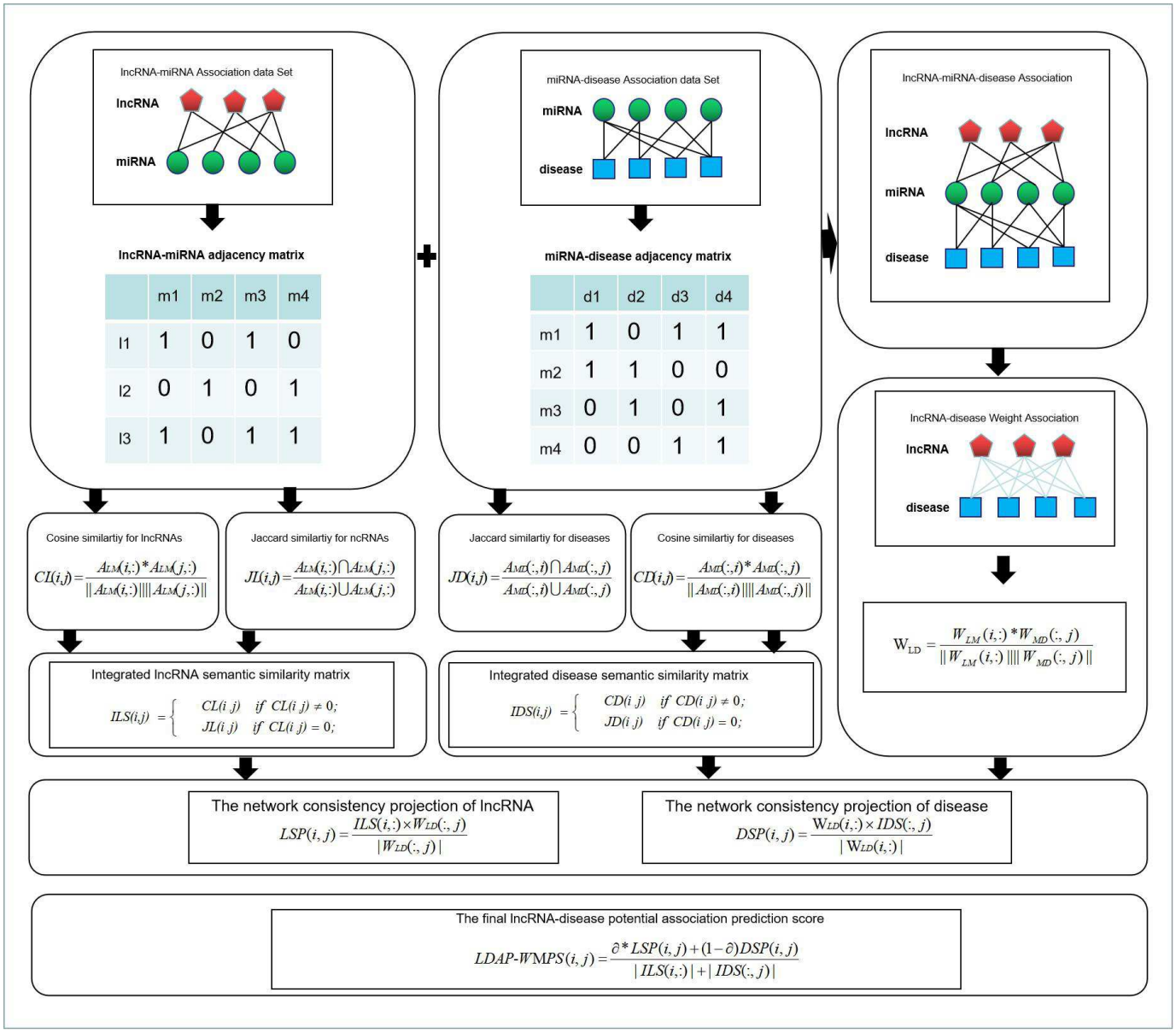

Figure 7. Flow Chart of FS-NCPLDA Applied to lncRNA-Disease Association Prediction

j; $A_{LM}(i,:) \cap A_{LM}(j,:)$ is the number of miRNAs associated with disease i and disease j, $A_{LM}(i,:) \cup A_{LM}(j,:)$ is the sum of miRNAs related to disease i and disease j

Integrated lncRNA similarity matrix

Integrated miRNA similarity MS and cosine similarity CL for lncRNA:

$$ILS(i,j) = \begin{cases} CL(i\ j) & if\ CL(i\ j) \neq 0; \\ JL(i\ j) & if\ CL(i\ j) = 0; \end{cases} \quad (8)$$

Establishment of lncRNA-disease weight matrix

Weight assignment algorithm [52] is often used in association prediction of lncRNA dual network. Through the weight distribution, we can get the correlation score between lncRNA-diseases. We further improved it and applied it to the lncRNA miRNA disease triple network, as shown in Figure 6. Taking L to M as an example, the first step is defined as:

$$f(M_j) = \sum_{i=1}^{m} \frac{a_{ij} f(L_i)}{k(L_i)} \quad (9)$$

Where m is the number of lncRNA types, $k(L_i)$ shows the number of miRNA species related to lncRNA i, and a_{ij} represents an entity in lncRNA-miRNA matrix A_{LM}.

The second step is M to D, defined as:

$$f(D_e) = \sum_{j=1}^{n} \frac{b_{je}}{k(M_j)} \sum_{i=1}^{m} \frac{a_{ij} f(L_i)}{k(L_i)} \quad (10)$$

Where n is the number of miRNA types, e is the number

of diseases types. $k(M_j)$ shows the number of diseases species related to miRNA j, and b_{je} represents an entity in miRNA-disease matrix A_{MD}.

And $f(D_e)$ can be expressed as:

$$f(D_e) = \sum_{i=1}^{m} s^{md} * f(L_i) \tag{11}$$

Combine (11) and (12) to get the formula：

$$s^{md} = \frac{1}{k(L_i)} \sum_{j=1}^{n} \frac{b_{je} a_{ij}}{k(M_j)} \tag{12}$$

In the above formula, $S_{MD} = \{s^{md}\}_{n*n}$ is the score of miRNA-disease association. miRNA disease association weight matrix is defined as:

$$W_{MD} = S_{MD} * A_{MD} \tag{13}$$

Similarly, the weight matrix W_{LM} from D to M to L is defined as:

$$s^{lm} = \frac{1}{k(D_e)} \sum_{j=1}^{n} \frac{b_{je} a_{ij}}{k(M_j)} \tag{14}$$

$$W_{LM} = A_{LM} * S_{LM} \tag{15}$$

For lncRNA i, we calculated the potential association characteristics between miRNAs related to lncRNA i, and for disease j, we also calculated the potential association characteristics between miRNAs related to disease j,We use $W_{LM}(i,:)$ to represent the eigenvalue of miRNA associated with lncRNA i and $W_{MD}(:,j)$ to represent the eigenvalue of miRNA associated with disease, Then the weight between lncRNA and disease is defined as:

$$w^{ij} = \frac{W_{LM}(i,:) * W_{MD}(:,j)}{\| W_{LM}(i,:) \| \| W_{MD}(:,j) \|} \tag{16}$$

Where $W_{LD} = \{w^{ij}\}$,The larger the value of w^{ij} , the larger the similarity of $W_{LM}(i,:)$ and $W_{MD}(:,j)$.

Building LDAP-WMPS Prediction Model

The flow chart of LDAP-WMPS model is shown in Figure 7. LDAP-WMPS model is divided into three parts, the first step is to calculate the disease projection score; the second step is to calculate the lncRNA projection score; the third step is to fuse the disease projection score and lncRNA projection score proportionally, and then normalize them to get our prediction score matrix.

The disease projection score is defined by the following formula:

$$DSP(i,j) = \frac{W_{LD}(i,:) \times IDS(:,j)}{| W_{LD}(i,:) |} \tag{17}$$

$W_{LD}(i,:)$ is the vector formed by the i row of lncRNA-disease weight matrix, which represents the association score between lncRNA i and various diseases. $IDS(:,j)$ is the vector formed by column j of the integrated disease similarity matrix, which represents the vector composed of the similarity between disease j and other diseases. $| W_{LD}(i,:) |$ represents module length of lncRNA i related disease component vector. $DSP(i,j)$ is the projection score of disease. The multi-dimensional similarity relation is transformed into concrete value by projection, the more similar lncRNA i is to lncRNA j, the higher the value of $DSP(i,j)$ is.

The projection score of lncRNA is defined as:

$$LSP(i,j) = \frac{ILS(i,:) \times W_{LD}(:,j)}{| W_{LD}(:,j) |} \tag{18}$$

In the above formula, $ILS(i,:)$ is the vector formed by the i-row of the functional similarity matrix of lncRNA, which represents the vector composed of the similarity between lncRNA i and other kinds of lncRNA. $W_{LD}(:,j)$ is the vector formed by column j of lncRNA-disease association weight matrix, which represents the association score between disease j and various lncRNAs. $| W_{LD}(:,j) |$ is the represents module length of lncRNA i related disease component vector. $LSP(i,j)$ is the projection score of lncRNA. Similarly, the more similar disease i is to disease j, the higher the value of $LSP(i,j)$ is.

The final lncRNA-disease potential association prediction score matrix was formed by fusing lncRNA projection score with disease projection score, defined as:

$$LDAP\text{-}WDPS(i,j)=\frac{\partial * LSP(i,j)+(1-\partial)DSP(i,j)}{|ILS(i,:)|+|IDS(:,j)|} \quad (19)$$

$LDAP\text{-}WDPS(i,j)$ is the final association score between lncRNA i and disease j. $|\mathrm{ILS}(i,:)|$ is the module length of lncRNA composition vector similar to lncRNA i in integrated lncRNA similarity matrix, and $|\mathrm{IDS}(:,j)|$ is the module length of disease composition vector similar to disease j in integrated disease similarity matrix. ∂ is the proportion of lncRNA projection score and disease projection score in fusion score calculation, We will analyze this parameter later.

Abbreviations

AUC: areas under ROC curve; LDAP-WMPS: lncRNA-Disease Association Prediction Based On Weight Distribution And Projection Score; PR: false positive rates; LMDN: the lncRNA-miRNA-disease tripartite network; LMDN: an updated lncRNA-miRNA-disease association tripartite network; lncRNA: long non-coding RNAs lncRNA; LOOCV: Leave-One Out Cross Validation; PR: true positive rates

Acknowledgments

The authors thank all those who have made suggestions for this article.

Author's contributions

WB conceived the study. WB, ZCdeveloped the method. DXX and ZJF implemented the algorithms. ZJF and WB collected the data. ZC performed the data analyses. WB and ZC wrote the manuscript. All authors have read and approved the manuscript

Funding

This work was supported in part by the grants of the Young Innovative Talents Project of Basic Scientific Research Business Expenses for Provincial Universities of Heilongjiang Province, No. 135509210.

Publication

Availability of data and materials

The Matlab code can be download at https://github.com/drbowang/LDAP-WMPS;

The datasets generated and/or analysed during the current study are available in the HMDD repository, http://www.cuilab.cn/;

MNDR repository, http://www.rna-society.org/mndr/;

starBase repository, http://starbase.sysu. edu.cn/starbase2/index.php;

Additional fle 1 Known miRNA-disease associations obtained from HMDD.

Additional fle 2 Known miRNA-lncRNA associations obtained from starBase v2.0.

Additional fle 3 Known lncRNA-disease associations obtained from MNDR v2.0.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that there are no competing interests regarding the publication of this paper.

Author details

B. Wang (Corresponding author) is with the College of Computer and Control, Qiqihar University, Qiqihar 161006, China (e-mail: drbowang@163.com). * Corresponding author.

C. Zhang is with the College of Computer and Control, Qiqihar University, Qiqihar 161006, China (e-mail: zc553070903@163.com).

X.-X Du is with the College of Computer and Control, Qiqihar University, Qiqihar 161006, China (e-mail: xiaoxin_du@163.com).

J.-F Zhang is with the College of Computer and Control, Qiqihar University, Qiqihar 161006, China (e-mail: jian_fei_zhang@163.com).

References

1. Mattick, John S., and Igor V. Makunin. "Non-Coding RNA." Human Molecular Genetics, vol. 15, 2006.

2. Gil, Noa, and Igor Ulitsky. "Regulation of Gene Expression by Cis-Acting Long Non-Coding RNAs." Nature Reviews Genetics, vol. 21, no. 2, 2020, pp. 102–117.

3. Yi, Kaining, et al. "Long Noncoding RNA and Its Role in Virus Infection and Pathogenesis." Frontiers in Bioscience, vol. 24, no. 4, 2019, pp. 777–789.

4. Chen, Linlin, et al. "lncRNA, MiRNA and lncRNA-MiRNA Interaction in Viral Infection." Virus Research, vol. 257, 2018, pp. 25–32.

5. Chen, Y.Grace, et al. "Gene Regulation in the Immune System by Long Noncoding RNAs." Nature Immunology, vol. 18, no. 9, 2017, pp. 962–972.

6. Guttman, Mitchell, et al. "Ab Initio Reconstruction of Cell Type–specific Transcriptomes in Mouse Reveals the Conserved Multi-Exonic Structure of LincRNAs." Nature Biotechnology, vol. 28, no. 5, 2010, pp. 503–510.

7. Hüttenhofer, Alexander, et al. "Non-Coding RNAs: Hope or Hype?" Trends in Genetics, vol. 21, no. 5, 2005, pp. 289–297.

8. Long, Yicheng, et al. "How Do lncRNAs Regulate Transcription." Science Advances, vol. 3, no. 9, 2017.

9. Ju, Cheng, et al. "Mesenchymal Stem Cell-Associated lncRNA in Osteogenic Differentiation." Biomedicine & Pharmacotherapy, vol. 115, 2019, p. 108912.

10. Quinn, Jeffrey J., and Howard Y. Chang. “Unique Features of Long Non-Coding RNA Biogenesis and Function.” Nature Reviews Genetics, vol. 17, no. 1, 2016, pp. 47 – 62.

11. Zhao, Wenyan, et al. “lncRNA HOTAIR Influences Cell Growth, Migration, Invasion, and Apoptosis via the MiR-20a-5p/HMGA2 Axis in Breast Cancer.” Cancer Medicine, vol. 7, no. 3, 2018, pp. 842 – 855.

12. Ferrè, Fabrizio, et al. “Revealing Protein – lncRNA Interaction.” Briefings in Bioinformatics, vol. 17, no. 1, 2016, pp. 106 – 116.

13. Bhan, Arunoday, et al. “Long Noncoding RNA and Cancer: A New Paradigm.” Cancer Research, vol. 77, no. 15, 2017, pp. 3965 – 3981.

14. Fernando, Thilini R., et al. “The lncRNA CASC15 Regulates SOX4 Expression in RUNX1-Rearranged Acute Leukemia.” Molecular Cancer, vol. 16, no. 1, 2017, pp. 126 – 126.

15. Delás, M.Joaquina, et al. “lncRNA Requirements for Mouse Acute Myeloid Leukemia and Normal Differentiation.” ELife, vol. 6, 2017.

16. Feng, Shui-Dong, et al. “Potential Regulatory Mechanisms of lncRNA in Diabetes and Its Complications.” Biochemistry and Cell Biology, vol. 95, no. 3, 2017, pp. 361 – 367.

17. Suwal, Abhishek, et al. “NONRATT021972 Long-Noncoding RNA: A Promising lncRNA in Diabetes-Related Diseases.” International Journal of Medical Sciences, vol. 16, no. 6, 2019, pp. 902 – 908.

18. Hua, Junjie Tony, et al. “Risk SNP-Mediated Promoter-Enhancer Switching Drives Prostate Cancer through lncRNA PCAT19.” Cell, vol. 174, no. 3, 2018, pp. 564 – 575.

19. Wu, Meng, et al. “lncRNA MEG3 Inhibits the Progression of Prostate Cancer by Modulating MiR-9-5p/QKI-5 Axis.” Journal of Cellular and Molecular Medicine, vol. 23, no. 1, 2019, pp. 29 – 38.

20. Loewen, Gregory, et al. “Functions of lncRNA HOTAIR in Lung Cancer.” Journal of Hematology & Oncology, vol. 7, no. 1, 2014, pp. 90 – 90.

21. Yx, Zhang, et al. “lncRNA TUC338 Promotes Invasion of Lung Cancer by Activating MAPK Pathway.” European Review for Medical and Pharmacological Sciences, vol. 22, no. 2, 2018, pp. 443 – 449.

22. Huang, Jin-Zhou, et al. “A Peptide Encoded by a Putative lncRNA HOXB-AS3 Suppresses Colon Cancer Growth.” Molecular Cell, vol. 68, no. 1, 2017, p. 171.

23. Wu, Qiong, et al. “lncRNA MALAT1 Induces Colon Cancer Development by Regulating MiR‐129‐5p/HMGB1 Axis.” Journal of Cellular Physiology, vol. 233, no. 9, 2018, pp. 6750 – 6757.

24. Huang, Ying. “The Novel Regulatory Role of lncRNA‐miRNA‐mRNA Axis in Cardiovascular Diseases.” Journal of Cellular and Molecular Medicine, vol. 22, no. 12, 2018, pp. 5768 – 5775.

25. Bär, Christian, et al. “Long Noncoding RNAs in Cardiovascular Pathology, Diagnosis, and Therapy.” Circulation, vol. 134, no. 19, 2016, pp. 1484 – 1499.

26. Huarte, Maite. “The Emerging Role of lncRNAs in Cancer.” Nature Medicine, vol. 21, no. 11, 2015, pp. 1253 – 1261.

27. Li, Jing-Jing, et al. “Long Non-Coding RNAs and Complex Human Diseases.” International Journal of Molecular Sciences, vol. 14, no. 9, 2013, pp. 18790 – 18808.

28. Li, Guanghui, et al. “Predicting MicroRNA-Disease Associations Using Label Propagation Based on Linear Neighborhood Similarity.” Journal of Biomedical Informatics, vol. 82, 2018, pp. 169 – 177.

29. Gu, Changlong, et al. “Network Consistency Projection for Human MiRNA-Disease Associations Inference.” Scientific Reports, vol. 6, no. 1, 2016, p. 36054.

30. Li, Guanghui, et al. “Prediction of lncRNA-Disease Associations Based on Network Consistency Projection.” IEEE Access, vol. 7, 2019, pp. 58849 – 58856.

31. Yu, Jingwen, et al. “A Novel Probability Model for lncRNA – Disease Association Prediction Based on the Naïve Bayesian Classifier.” Genes, vol. 9, no. 7, 2018, p. 345.

32. Chen, Xing, and Gui-Ying Yan. “Novel Human lncRNA-Disease Association Inference Based on lncRNA Expression Profiles.” Bioinformatics, vol. 29, no. 20, 2013, pp. 2617 – 2624.

33. Yu, Jingwen, et al. “A Novel Collaborative Filtering Model for lncRNA-Disease Association Prediction Based on the Naïve Bayesian Classifier.” BMC Bioinformatics, vol. 20, no. 1, 2019, pp. 1 – 13.

34. Lu, Chengqian, et al. “Prediction of lncRNA-Disease Associations Based on Inductive Matrix Completion.” Bioinformatics, vol. 34, no. 19, 2018, pp. 3357 – 3364.

35. Dongfeng, Yin. “Experience of Professor YIN Dongfeng′s Treatment on Oncology Under the Guidance of TCM Theory.” Liaoning Journal of Traditional Chinese Medicine, 2013.

36. Brambilla, E., et al. “The New World Health Organization Classification of Lung Tumours.” European Respiratory Journal, vol. 18, no. 6, 2001, pp. 1059 – 1068.

37. Chen, Dong Liang, et al. “Long Noncoding RNA XIST Expedites Metastasis and Modulates Epithelial-Mesenchymal Transition in Colorectal Cancer.” Cell Death and Disease, vol. 8, no. 8, 2017.

38. Chen, Zhi-Yuan, et al. “lncRNA SNHG16 Promotes Colorectal Cancer Cell Proliferation, Migration, and Epithelial-Mesenchymal Transition through MiR-124-3p/MCP-1.” Gene Therapy, 2020, pp. 1 – 13.

39. Wu, Chuanqing, et al. “MALAT1 Promotes the Colorectal Cancer Malignancy by Increasing DCP1A Expression and MiR203 Downregulation.” Molecular Carcinogenesis, vol. 57, no. 10, 2018, pp. 1421 – 1431.

40. Li, Shunle, et al. “The Long Non‐coding RNA HCG18 Promotes the Growth and Invasion of Colorectal Cancer Cells through Sponging MiR‐1271 and Upregulating MTDH/Wnt/β‐catenin.” Clinical and Experimental Pharmacology and Physiology, vol. 47, no. 4, 2020, pp. 703 – 712.

41. Li, Ding, et al. “Long Noncoding RNA FGD5-AS1 Promotes Colorectal Cancer Cell Proliferation, Migration, and Invasion through Upregulating CDCA7 via Sponging MiR-302e.” In Vitro Cellular & Developmental Biology – Animal, vol. 55, no. 8, 2019, pp. 577 – 585.

42. Wang, Meng, et al. “Long Non-Coding RNA TUG1 Mediates 5-Fluorouracil Resistance by Acting as a CeRNA of MiR-197-3p in Colorectal Cancer.” Journal of Cancer, vol. 10, no. 19, 2019, pp. 4603 – 4613.

43. Kuhn, E., et al. “Adenocarcinoma Classification: Patterns and Prognosis.” Pathologica, vol. 110, no. 1, 2018, pp. 5 – 11.

44. Sun, Jing, et al. “lncRNA XIST Promotes Human Lung Adenocarcinoma Cells to Cisplatin Resistance via Let-7i/BAG-1 Axis.” Cell Cycle, vol. 16, no. 21, 2017, pp. 2100–2107.

45. Lu, Zhengmao, et al. “MALAT1 Promotes Gastric Adenocarcinoma through the MALAT1/MiR-181a-5p/AKT3 Axis.” Open Biology, vol. 9, no. 9, 2019, p. 190095.

46. Du, Zhou, et al. “Integrative Analyses Reveal a Long Noncoding RNA-Mediated Sponge Regulatory Network in Prostate Cancer.” Nature Communications, vol. 7, no. 1, 2016, pp. 10982–10982.

47. Li, Wei, et al. “HCG18/MiR-34a-5p/HMMR Axis Accelerates the Progression of Lung Adenocarcinoma.” Biomedicine & Pharmacotherapy, vol. 129, 2020, p. 110217.

48. Guo, Yuxia, et al. “Long Non-Coding RNA SNHG16 Promotes Cell Proliferation and Invasion in Lung Adenocarcinoma via Sponging Let-7a-5p.” Minerva Chirurgica, vol. 74, no. 6, 2020, pp. 509–511.

49. Li, Yang, et al. “HMDD v2.0: A Database for Experimentally Supported Human MicroRNA and Disease Associations.” Nucleic Acids Research, vol. 42, 2014, pp. 1070–1074.

50. Li, Jun Hao, et al. “StarBase v2.0: Decoding MiRNA-CeRNA, MiRNA-NcRNA and Protein–RNA Interaction Networks from Large-Scale CLIP-Seq Data.” Nucleic Acids Research, vol. 42, 2014, pp. 92–97.

51. Cui, Tianyu, et al. “MNDR v2.0: An Updated Resource of NcRNA–disease Associations in Mammals.” Nucleic Acids Research, vol. 46, 2017, pp. 371–374.

52. Yang, Xiaofei, et al. “A Network Based Method for Analysis of lncRNA-Disease Associations and Prediction of lncRNAs Implicated in Diseases.” PLOS ONE, vol. 9, no. 1, 2014.

Figures

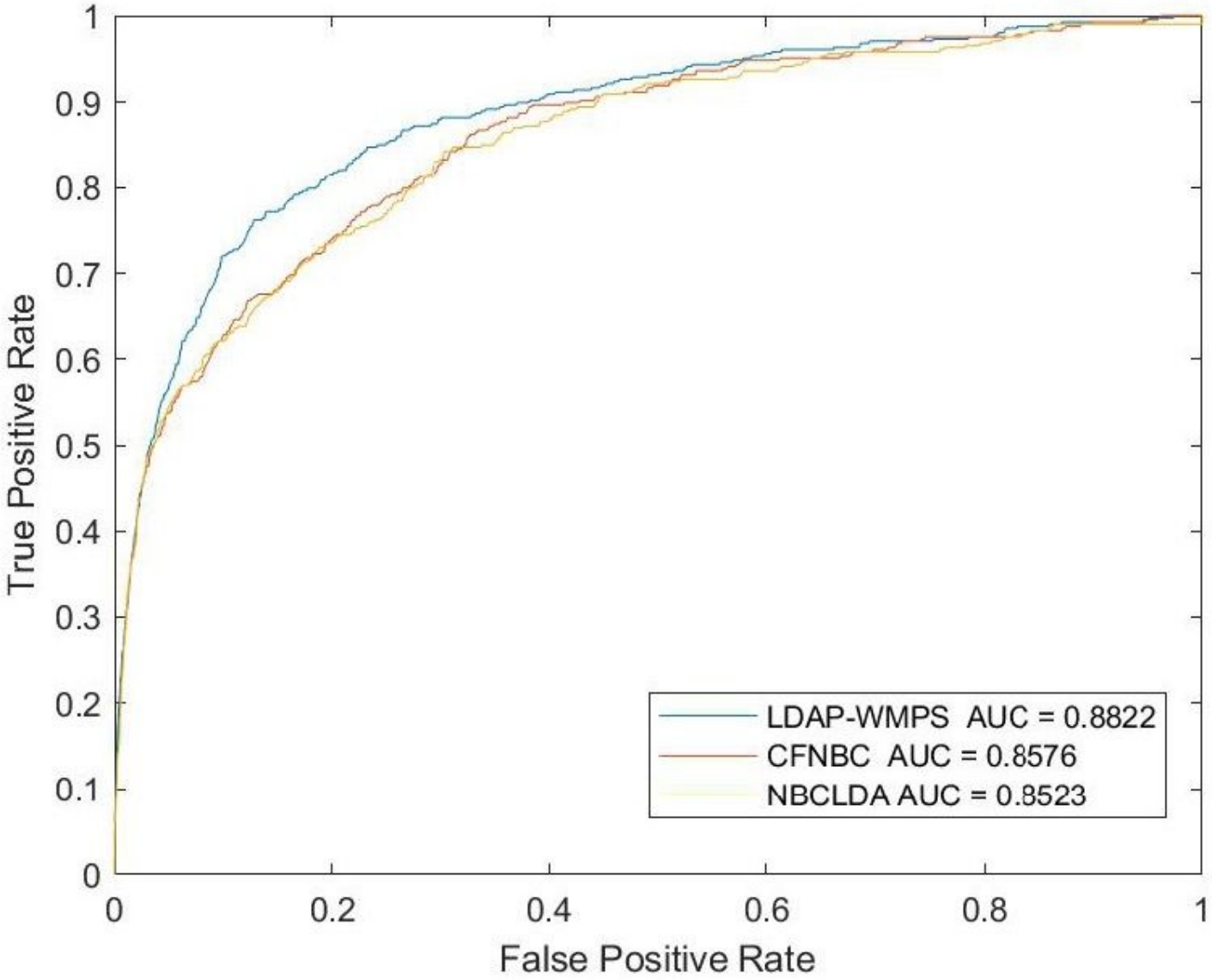

Figure 1

The performance of LDAP-WMPS and others models in terms of ROC curves and AUCs based on 407 known lncRNA-disease associations under the framework of LOOCV

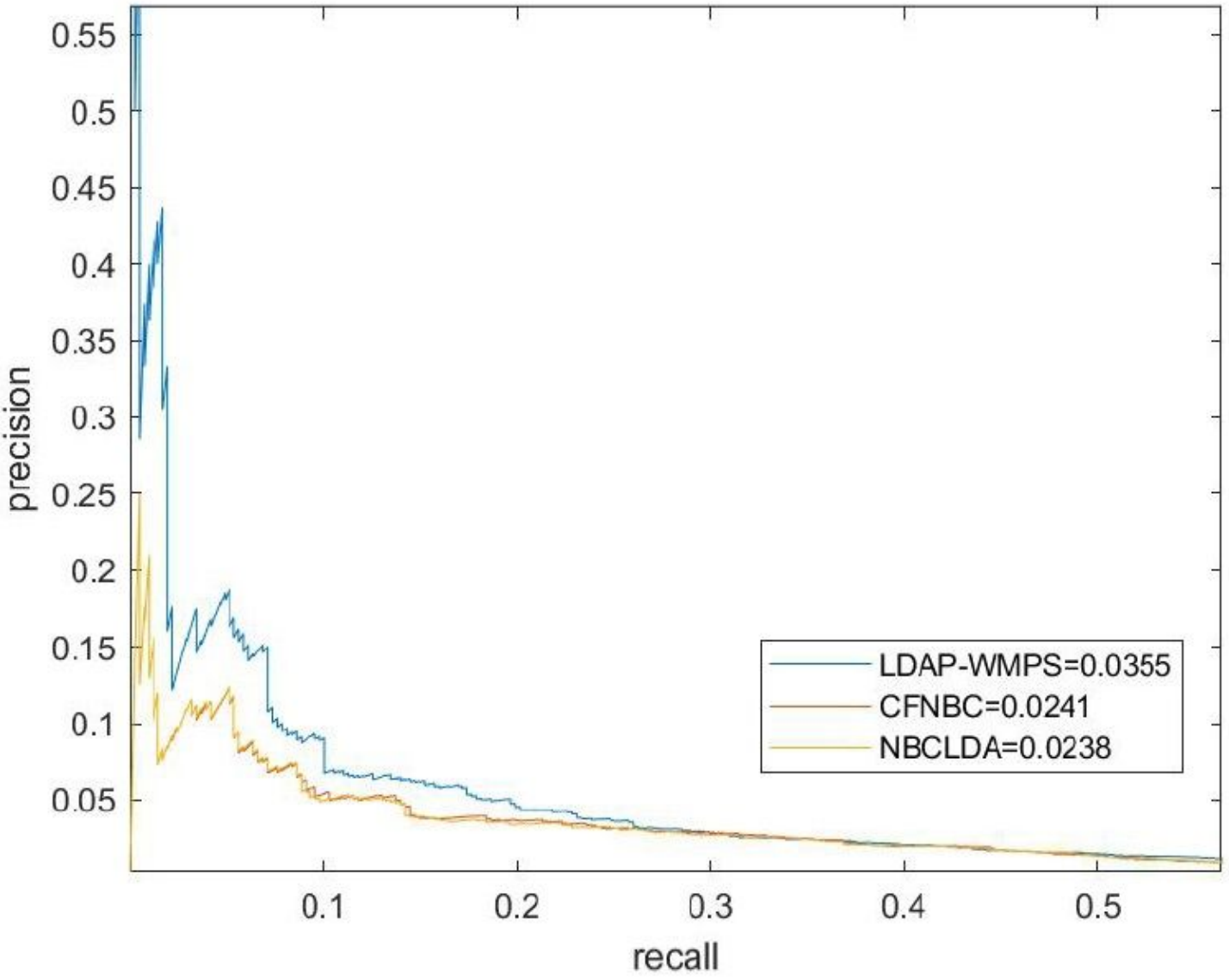

Figure 2

The performance of LDAP-WMPS and others models in terms of PR curves and AUPRs based on 407 known lncRNA-disease associations under the framework of LOOCV

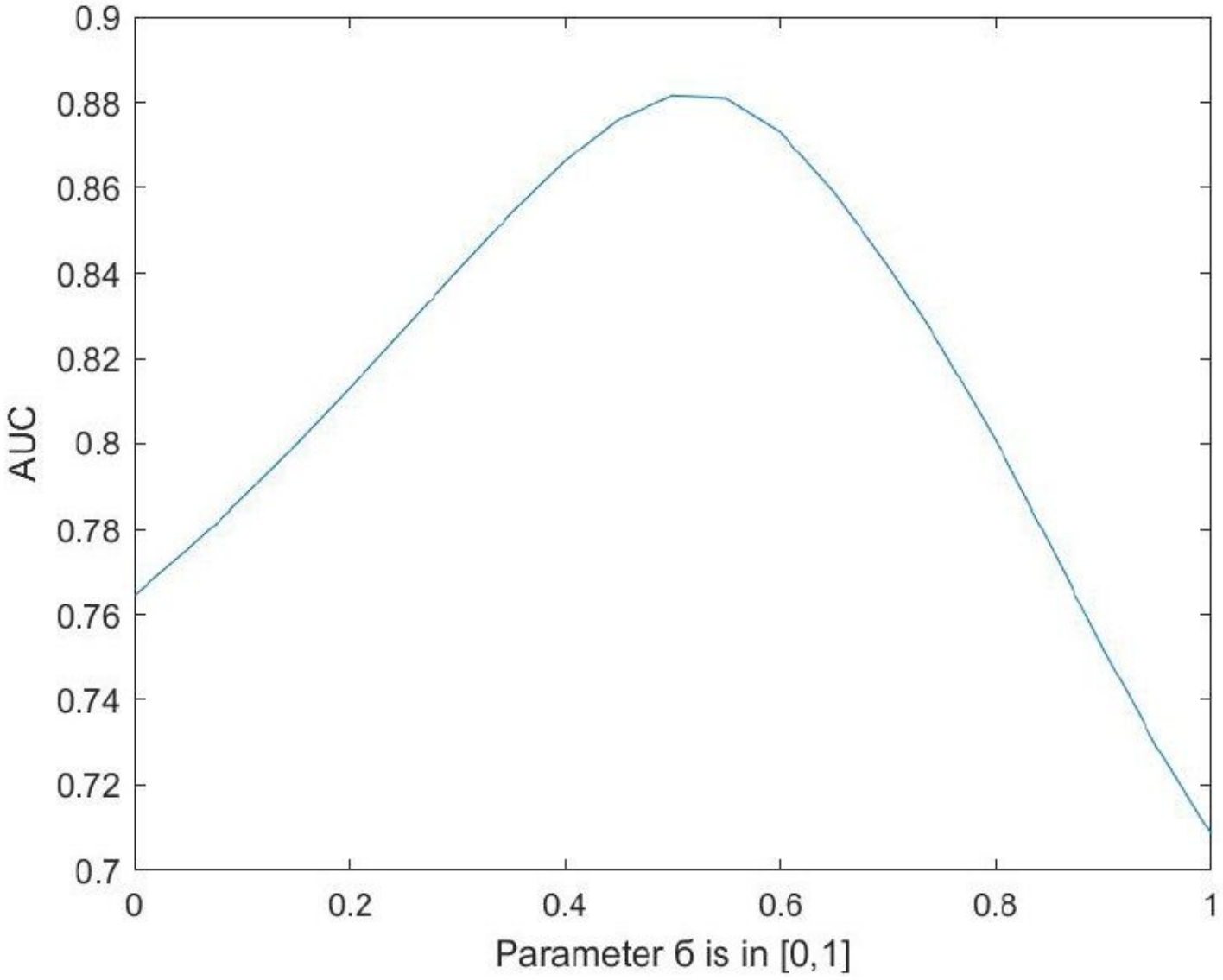

Figure 3

Transformation curve of parameter in the range of [0,1]

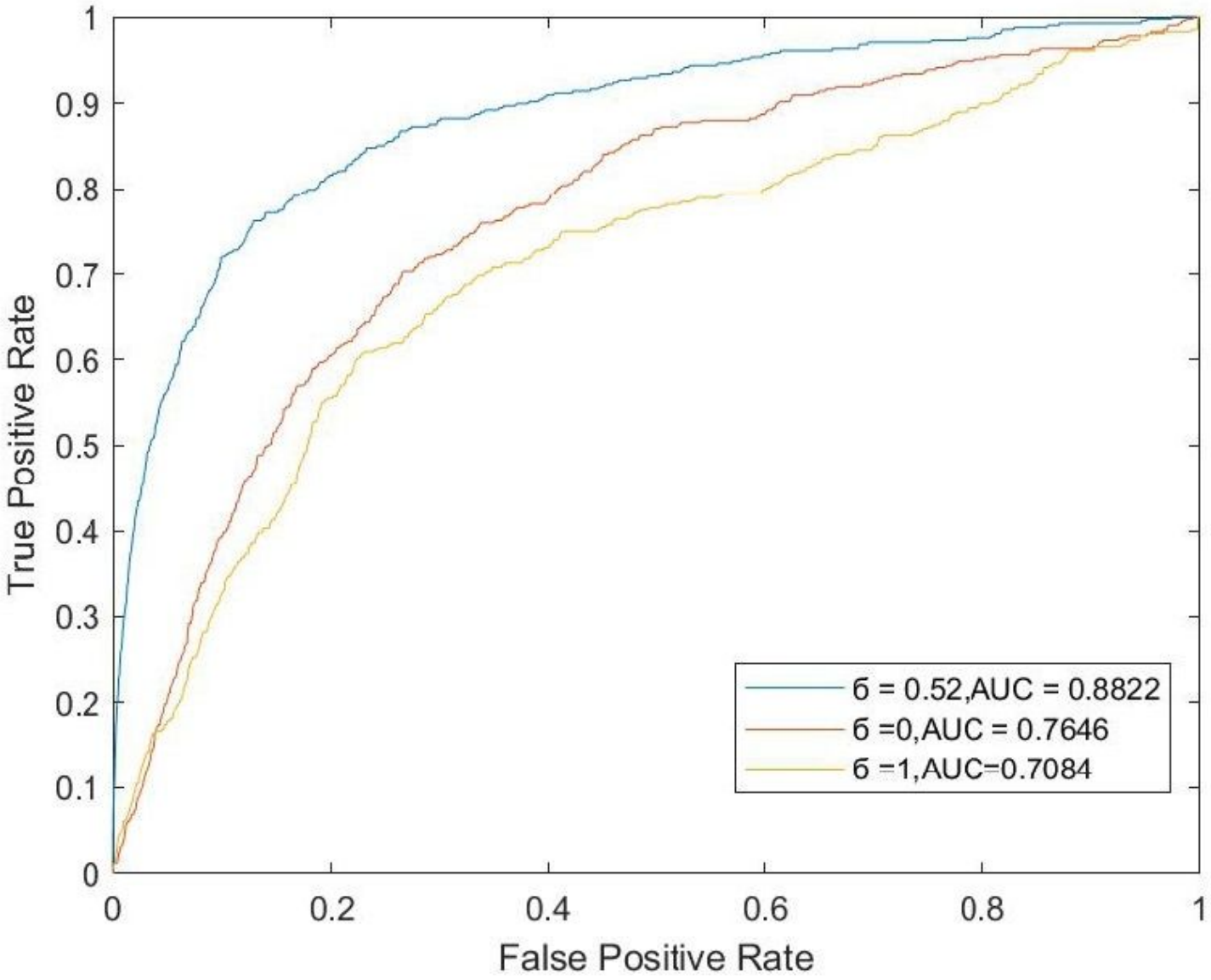

Figure 4

ROC calculated by fusion of lncRNA projection fraction and disease projection fraction was compared with ROC calculated by lncRNA projection fraction only and disease projection fraction only.

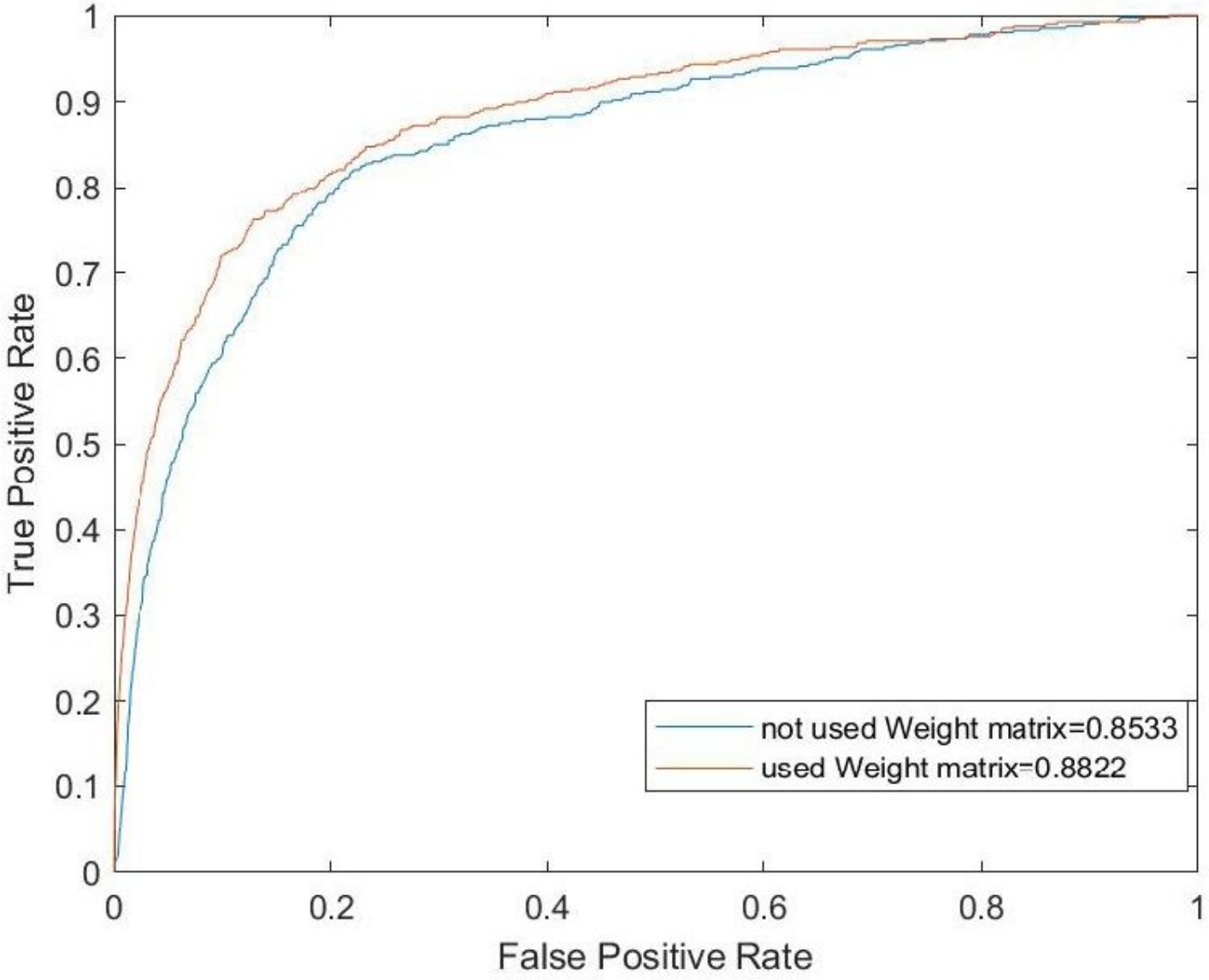

Figure 5

Comparison of ROC curve calculated with weight matrix and ROC curve calculated without weight matrix.

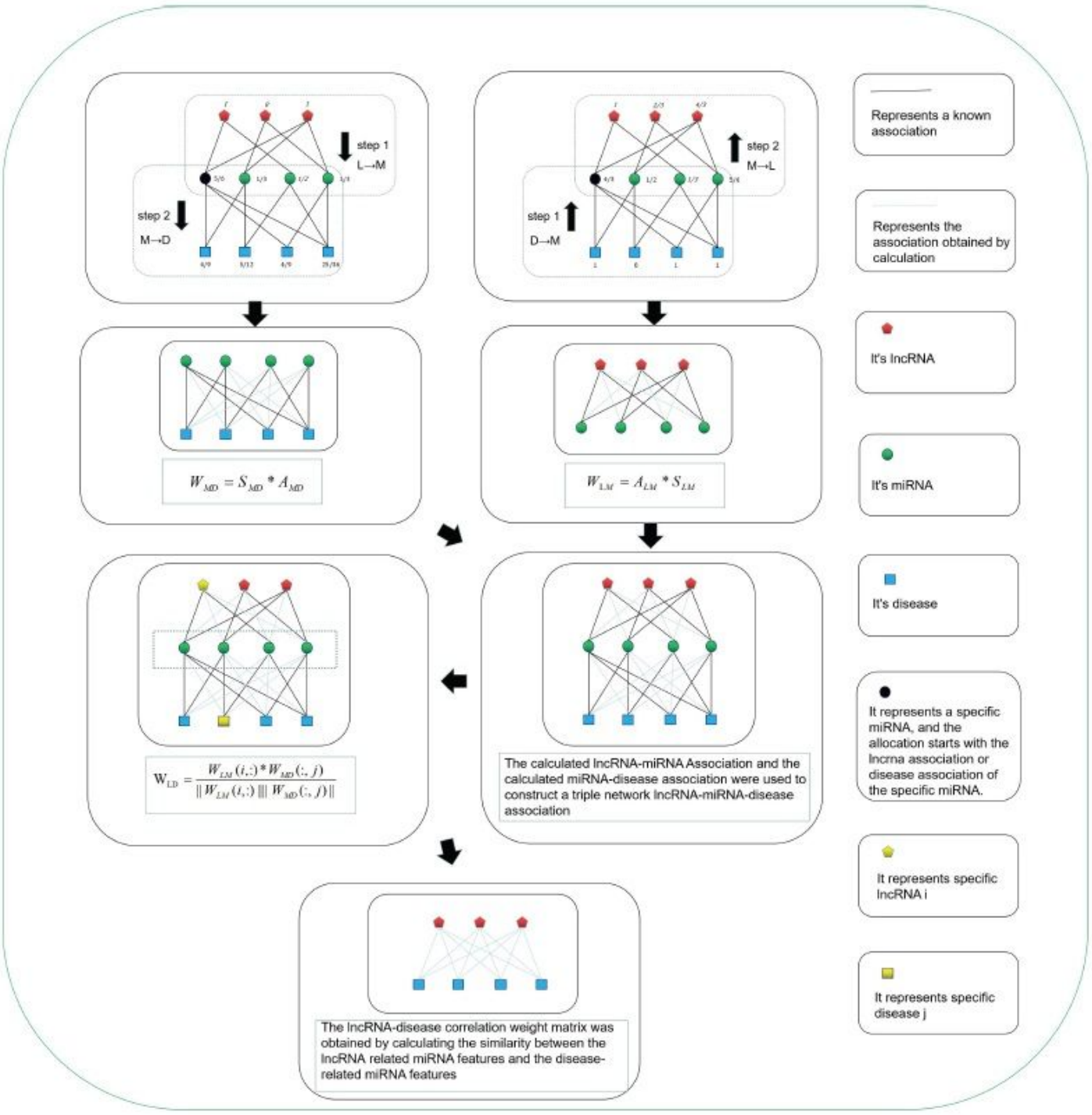

Figure 6

Flow chart of lnRNA-disease association weight matrix construction

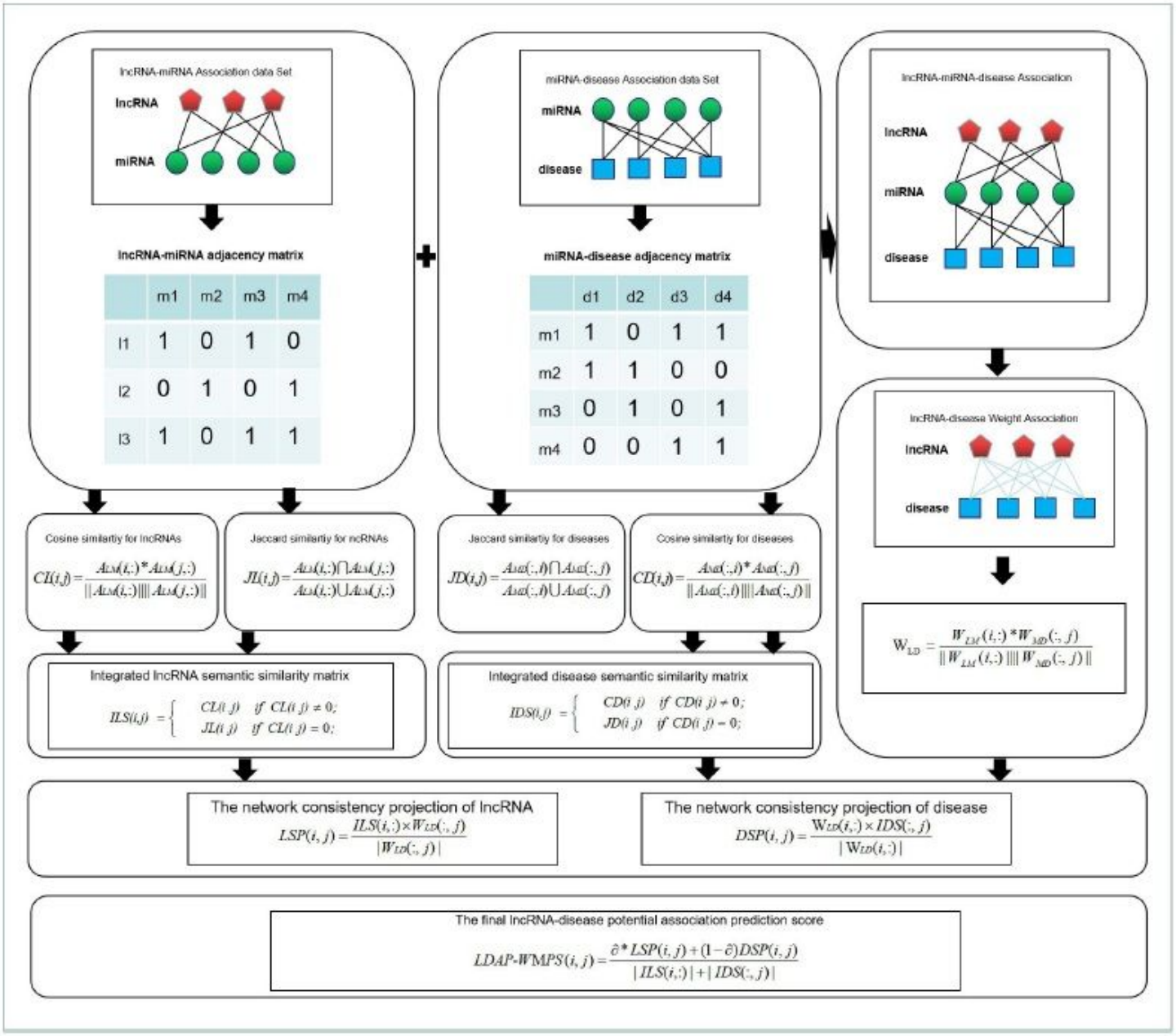

Figure 7

Flow Chart of FS-NCPLDA Applied to lncRNA-Disease Association Prediction

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- KnownmiRNAdiseaseassociationsobtainedfromHMDD.xlsx
- KnownmiRNAlncRNAassociationsobtainedfromstarBasev2.0.xlsx
- KnownlncRNAdiseaseassociationsobtainedfromMNDRv2.0.xlsx
- KnownlncRNAdiseaseassociationsobtainedfromMNDRv2.0.xlsx

- KnownmiRNAdiseaseassociationsobtainedfromHMDD.xlsx
- KnownmiRNAlncRNAassociationsobtainedfromstarBasev2.0.xlsx