Bioinformatics, 2018, 1–9
doi: 10.1093/bioinformatics/bty333
Advance Access Publication Date: 25 April 2018
Original Paper

OXFORD

Systems biology

BNPMDA: Bipartite Network Projection for MiRNA–Disease Association prediction

Xing Chen[1,*], Di Xie[2], Lei Wang[1], Qi Zhao[2,3,*], Zhu-Hong You[4] and Hongsheng Liu[3,5]

[1]School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China, [2]School of Mathematics, Liaoning University, Shenyang 110036, China, [3]Research Center for Computer Simulating and Information Processing of Bio-Macromolecules of Liaoning Province, Shenyang 110036, China, [4]Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Science, Ürümqi 830011, China and [5]School of Life Science, Liaoning University, Shenyang 110036, China

*To whom correspondence should be addressed.
Associate Editor: Alfonso Valencia

Received on March 5, 2017; revised on March 19, 2018; editorial decision on April 23, 2018; accepted on April 24, 2018

Abstract

Motivation: A large number of resources have been devoted to exploring the associations between microRNAs (miRNAs) and diseases in the recent years. However, the experimental methods are expensive and time-consuming. Therefore, the computational methods to predict potential miRNA–disease associations have been paid increasing attention.

Results: In this paper, we proposed a novel computational model of Bipartite Network Projection for MiRNA–Disease Association prediction (BNPMDA) based on the known miRNA–disease associations, integrated miRNA similarity and integrated disease similarity. We firstly described the preference degree of a miRNA for its related disease and the preference degree of a disease for its related miRNA with the bias ratings. We constructed bias ratings for miRNAs and diseases by using agglomerative hierarchical clustering according to the three types of networks. Then, we implemented the bipartite network recommendation algorithm to predict the potential miRNA–disease associations by assigning transfer weights to resource allocation links between miRNAs and diseases based on the bias ratings. BNPMDA had been shown to improve the prediction accuracy in comparison with previous models according to the area under the receiver operating characteristics (ROC) curve (AUC) results of three typical cross validations. As a result, the AUCs of Global LOOCV, Local LOOCV and 5-fold cross validation obtained by implementing BNPMDA were 0.9028, 0.8380 and 0.8980 $\pm$ 0.0013, respectively. We further implemented two types of case studies on several important human complex diseases to confirm the effectiveness of BNPMDA. In conclusion, BNPMDA could effectively predict the potential miRNA–disease associations at a high accuracy level.

Availability and implementation: BNPMDA is available via http://www.escience.cn/system/file?fileId=99559.

Contact: xingchen@amss.ac.cn or zhaoqi.shenyang@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The first microRNA (miRNA) named lin-4 was found twenty years ago by Victor Ambros (Lee *et al.*, 1993). Soon after, a large amount of miRNAs have been discovered from a wide variety of species (Jopling *et al.*, 2005). Studies have uncovered that miRNAs are non-coding RNAs including approximately 22 nucleotides (Ribeiro

© The Author(s) 2018. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

Downloaded from https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/bty333/4986418
by Xiamen University user
on 07 June 2018

et al., 2014). In general, they could bind specific target messenger RNAs through base-pairing interactions and repress their translation and stability. Besides, the expression pattern of a miRNA could provide important clues toward its function (Pasquinelli, 2016). The miRNAs could influence some important biological processes, such as cell development, proliferation and apoptosis (Bartel, 2009). Meanwhile, their regulatory functions are related to some special gene expressions in the post transcription stage (Wightman *et al.*, 1993). In recent researches, many experiments have shown that large amounts of miRNAs are associated with the development processes of various human diseases (Meola *et al.*, 2009). For example, the study of Kalinowski *et al.* showed that miR-7 could regulate epidermal growth factor receptor (EGFR) expression and protein kinase B (Akt) activity in head and neck cancer (HNC) cell lines (Kalinowski *et al.*, 2012). In 2010, the study of Amit *et al.* found that a number of miRNAs such as miR-193a, miR-224/miR-452 cluster, miR-182/miR-183/miR-96 cluster and miR-148a having potential tumor/metastasis suppressive activity were over-expressed in the WNT signaling associated medulloblastomas (Gokhale *et al.*, 2010). Furthermore, Von *et al.* had shown that miR-15a which was inversely correlated to protein kinase C (PKC) was a potential marker to differentiate between benign and malignant renal tumors in biopsy and urine samples (Von *et al.*, 2012). It is obvious to conclude that miRNAs have tight associations with human diseases. However, the miRNA–disease associations identified by experimental methods nowadays are only the tip of the iceberg. Furthermore, experimental methods for finding associations between miRNAs and diseases are expensive and time-consuming. Therefore, increasing numbers of studies have paid attention to the computational algorithms for predicting the potential miRNA–disease associations. According to the results of the prediction, the biological experiments could be implemented effectively by selecting the most promising disease-associated miRNAs (Bandyopadhyay *et al.*, 2010).

Based on the assumptions that miRNAs which have similar functions are more likely associated with similar disease and vice versa, Jiang *et al.* (2010) proposed a computational model by using hypergeometric distribution to identify the potential miRNA–disease s associations. However, it adopted the similarity information just by judging whether the pairs were similar, rather than using similarity scores. Another shortage was that they only used the local similarity of two miRNAs that had significant number of shared target genes. Li *et al.* (2011) proposed a computational method for predicting the miRNA–disease associations by calculating the function consistence score (FCS) between the target genes of miRNA and the disease-related genes. However, when calculating the FCS, this method ignored the topological structure of network composed of the targets and disease genes. Shi *et al.* (2013) developed a modified random walk algorithm by using the miRNA–target interactions, disease-gene associations and protein–protein interactions (PPIs). Mørk *et al.* (2014) presented a computational method called miRPD, which explicitly took advantage of the protein link between miRNA and disease that connected miRNA–protein interactions and protein–disease associations. The scoring function was constructed by multiplying the score of protein–disease association with the score of protein-miRNA interaction. Besides linking miRNAs to diseases, it had directly suggested the underlying proteins involved. Xu *et al.* (2014) constructed a miRNA prioritization method by integrating known disease–gene associations and miRNA–target interactions to prioritize novel disease-related miRNAs. Instead of using the known miRNA–disease associations, they needed to evaluate the similarity between the targets of miRNAs and disease genes. For all the aforementioned methods, because the miRNA–target interactions have high false positive and false negative samples, so they could not generate sufficiently accurate prediction results.

To deal with the limitations above, Xuan *et al.* (2013) constructed an HDMP model to predict disease-related miRNAs. When calculating miRNA similarity, they assigned higher weights to the similarity scores between two miRNAs associated with the given disease in the same family or cluster through multiplying similarity scores by the weight which was a function about the proportion of disease related miRNAs in the same family or cluster. However, HDMP could not predict the association between new diseases and miRNAs. Simultaneously, it was not suitable for diseases with a few known miRNAs. HDMP also could not perform better than most of previous models which were calculated based on the global network similarity measures. Chen *et al.* (2012) found that global network similarity is more effective to capture the associations between diseases and miRNAs than traditional local network similarity. Therefore, the method of RWRMDA had been developed to infer potential miRNA–disease associations by implementing random walk (a global network similarity measure-based algorithm). Chen *et al.* (2016c) presented another model of WBSMDA based on calculating the Gaussian interaction profile kernel similarity together with miRNA functional similarity and disease semantic similarity. WBSMDA could predict the potential related miRNAs of new diseases and new miRNAs without known association information. Chen *et al.* (2016a) also presented the computational model of HGIMDA to predict new disease-related miRNAs by integrating known miRNA–disease associations and different types of disease similarity and miRNA similarity into a heterogeneous graph. HGIMDA could find the optimal solutions with an iterative process based on global network similarity information. HGIMDA had an improved performance in comparison with the previous local network-similarity-based models for predicting miRNA–disease associations.

Nowadays, the machine learning algorithms such as Support Vector Machine (SVM) classifiers and some semi-supervised learning models have been applied in the bioinformatics and computational biology (Wong *et al.*, 2015). As an instance, Xu *et al.* (2011) introduced an approach based on the miRNA target-dysregulated network (MTDN) to prioritize novel disease-related miRNAs. They constructed the SVM classifier to distinguish positive miRNA–disease associations from negative ones by extracting the feature of network topologic information. However, as is known, it is hard to obtain the negative miRNA–disease associations. Thus, the ambiguity caused by negative samples influences the accuracy of the supervised classifier. Chen and Yan (2015) constructed a RLSMDA model based on semi-supervised learning for predicting potential disease-related miRNAs. RLSMDA could calculate prediction score of miRNA–disease association for new disease. Meanwhile, RLSMDA could avoid using negative associations between miRNAs and diseases. However, the parameter choice of RLSMDA and the ways of combining the classifiers in different spaces together need to be studied furthermore. In addition, Chen *et al.* (2015) developed the RBMMMDA method based on the restricted Boltzmann machine (RBM) with a two-layer (visible and hidden) undirected graphical miRNA–disease associations. Different from the previous methods, RBMMMDA could obtain the types of new miRNA–disease associations.

In this paper, we proposed the Bipartite Network Projection for MiRNA–Disease Association prediction (BNPMDA) model based on the rating-integrated bipartite network recommendation and the known miRNA–disease associations. Firstly, based on the miRNA functional similarity and the disease semantic similarity, the bias ratings between diseases and miRNAs were constructed by using agglomerative hierarchical clustering. Furthermore, the baseline algorithm for personal recommendation based on bipartite network projection was improved with the constructed bias ratings for a

Downloaded from https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/bty333/4986418
by Xiamen University user
on 07 June 2018

more accurate prediction of potential miRNA–disease associations. For estimating the prediction accuracy of BNPMDA, we implemented the global and local leave-one-out cross validation (LOOCV) and obtained the AUCs of 0.9028 and 0.8380, respectively. Furthermore, the 5-fold cross validation showed an average AUC of 0.8980 ± 0.0013. In order to further validation, we also carried out the case studies on the recent version of HMDD database and the previous version of HMDD database, respectively. As a result, there were high proportions of the predicted miRNAs confirmed by recent experimental reports. Therefore, we could conclude that BNPMDA has been confirmed to be powerful and effective in predicting potential miRNA–disease associations.

2 Materials and methods

2.1 Human miRNA–disease associations

The known human miRNA–disease associations data were downloaded from HMDD v2.0 database (Li *et al.*, 2014) which contains 5430 distinct experimentally confirmed human miRNA–disease s associations about 383 diseases and 495 miRNAs. For convenience, we constructed an adjacency matrix $A \in R^{nd \times nm}$ to formalize the human miRNA–disease associations, where *nm* and *nd* were denoted as the number of known miRNAs and known diseases. If a disease d_i had been experimentally verified to be associated with a miRNA m_j, then A_{ij} equals to 1, otherwise 0.

2.2 MiRNA functional similarity

Based on the assumption that functional similar miRNAs are more likely to be associated with similar diseases and vice versa, Wang *et al.* (2010) proposed method to calculate the miRNA functional similarity (Lu *et al.*, 2008). We could download the miRNA functional similarity data conveniently from http://www.cuilab.cn/files/images/cuilab/misim.zip owing to their excellent work. We denoted the matrix *FS* to represent the miRNA functional similarity. The element $FS(m_i, m_j)$ represents the value of similarity between the miRNA m_i and the miRNA m_j.

2.3 Disease semantic similarity model 1

A Directed Acyclic Graph (DAG) was constructed to describe the diseases according to the literature of Wang *et al.* (2010) based on the Medical Subject Headings (MeSH) descriptors which could be downloaded from the National Library of Medicine (http://www.nlm.nih.gov/). We provided the Supplementary Figure S1 as an example to explain the DAG based on MeSH (see Supplementary Fig. S1). According to the DAG, we denoted the contribution values of disease d in DAG(D) to the semantic value of disease D as follows:

$$\begin{cases} D1_D(d) = 1 \quad if \quad d = D \\ D1_D(d) = \max\{\Delta * D1_D(d') | d' \in children\ of\ d\} \quad if \quad d \neq D \end{cases} \tag{1}$$

where Δ was the semantic contribution decay factor. The semantic value of disease D was defined as follows:

$$DV1(D) = \sum_{d \in T(D)} D1_D(d) \tag{2}$$

where $T(D)$ represented all ancestor nodes of D and D itself. According to the observation that two diseases with larger shared part of their DAGs tend have larger similarity score, the semantic similarity score between disease d_i and d_j could be defined as follows:

$$SS1(d_i, d_j) = \frac{\sum_{t \in T(d_i) \cap T(d_j)} \left(D1_{d_i}(t) + D1_{d_i}(t) \right)}{DV1(d_i) + DV1(d_j)} \tag{3}$$

2.4 Disease semantic similarity model 2

As what we had considered, it was not reasonable to assign the same contribution value to the diseases in the same layer of DAG(D). Specially, a more specific disease that appears in less DAGs should contribute a higher value to the semantic similarity of disease D. Therefore, according to the model which was proposed by Xuan *et al.* (2013), we defined the contribution of disease d in DAG(D) to the semantic value of disease D as follows:

$$D2_D(d) = -\log[\text{the number of DAGs including } d / \text{the number of diseases}] \tag{4}$$

where d represented any disease of all the diseases under investigation. We defined the semantic similarity between disease d_i and d_j as the ratio of contributions from the shared ancestor nodes to the contributions from all the ancestor nodes. Consequently, the disease semantic similarity could be calculated as follows:

$$SS2(d_i, d_j) = \frac{\sum_{t \in T(d_i) \cap T(d_j)} \left(D2_{d_i}(t) + D2_{d_i}(t) \right)}{DV2(d_i) + DV2(d_j)} \tag{5}$$

where

$$DV2(D) = \sum_{d \in T(D)} D2_D(d) \tag{6}$$

2.5 Gaussian interaction profile kernel similarity

Based on the Gaussian kernel function which is one of the Radial Basis functions whose values depend only on the distance from the origin, Gaussian interaction profile kernel similarities were constructed as another algorithm of measuring disease similarity and miRNA similarity (Chen *et al.*, 2016b,c). Since the *i*th row and *j*th column of adjacent matrix A contains the information whether the disease or the miRNA associated with miRNA m_j or disease d_i, we denoted vector $IP(d_i)$ and $IP(m_j)$ to represent the *i*th row vector and the *j*th column vector for the further convenient utilization. Therefore, the similarity between diseases and miRNAs could be computed as follows:

$$GD(d_i, d_j) = \exp\left(-\beta_d \| IP(d_i) - IP(d_j) \|^2 \right) \tag{7}$$

$$GR(m_i, m_j) = \exp\left(-\beta_r \| IP(m_i) - IP(m_j) \|^2 \right) \tag{8}$$

where adjustment coefficient β_d and β_r for the kernel bandwidth were denoted as follows:

$$\beta_d = \beta'_d / \left(\frac{1}{nd} \sum_{i=1}^{n} \| IP(d_i) \|^2 \right) \tag{9}$$

$$\beta_m = \beta'_m / \left(\frac{1}{nm} \sum_{i=1}^{m} \| IP(m_i) \|^2 \right) \tag{10}$$

where β'_d and β'_m were the original bandwidths. In the end, matrix GD and GS were denoted to represent the Gaussian interaction profile kernel similarity of diseases and miRNAs, respectively.

Downloaded from https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/bty333/4986418
by Xiamen University user
on 07 June 2018

2.6 Integrated similarity for miRNAs and diseases

Based on the multiple similarity algorithms constructed above, the more accurate integrated disease similarity and integrated miRNA similarity were constructed by combining the Gaussian interaction profile kernel similarity with the disease semantic similarity and the miRNA functional similarity, respectively. For example, if disease d_i and d_j have semantic similarity, then the final integrated similarity equals to the average of *SS1* and *SS2*, otherwise it equals to the Gaussian interaction profile kernel similarity. The formula has been shown as follows:

$$SM(m_i, m_j) = \begin{cases} FS(m_i, m_j) & m_i \text{ and } m_j \text{ has functional similarity} \\ GR(m_i, m_j) & \text{otherwise} \end{cases} \tag{11}$$

$$SD(d_i, d_j) = \begin{cases} \dfrac{SS1(d_i, d_j) + SS2(d_i, d_j)}{2} & d_i \text{ and } d_j \text{ has semantic similarity} \\ GD(d_i, d_j) & \text{otherwise} \end{cases} \tag{12}$$

2.7 BNPMDA

All the subsections presented above from 2.1 to 2.6 are materials prepared for the following main algorithm to take advantage of. The six kinds of materials are basically coincident with two latest proposed models, namely WBSMDA (Chen *et al.*, 2016c) and HGIMDA (Chen *et al.*, 2016a). However, the earlier models had no such complete training data, such as the RLSMDA (Chen and Yan, 2015) which was not developed based on the integrated similarities for diseases and miRNAs due to the unconscious of Gaussian interaction profile kernel similarity. The earlier proposed model was RWRMDA, which was developed based on only human miRNA–disease associations and miRNA functional similarity. Moreover, RBMMMDA was proposed only based on human miRNA–disease associations due to the limitations of the algorithm. Therefore, all the materials used in our model are developed accumulatively over time. Nowadays, we use as many materials as possible to guarantee that we have made full use of the information of known data. Each type of the material appeared in history will be used unless a better version came out. The comparison figure has been shown as Supplementary Figure S2 which is the comparison diagram of BNPMDA and five previous models. In the first section of the diagram, the whole models are arranged with the time tendency. In the second section, we list the specific materials for each model. If one model has the given type of materials, the rectangular strip of the corresponding material will extend to the area under the model. In the third section, we list the core prediction algorithms of each model to show the main differences between all the models.

Through analyzing the known miRNA–disease associations, a given disease showed obviously different ratings of tendency to be associated with different miRNAs. For example, a given disease d_i may have association with many similar miRNAs, but some other d_i -related miRNAs do not have any similar miRNAs that are also associated with disease d_i. Therefore, based on the assumption that disease associated with a greater number of similar miRNAs implies a higher bias rating to these miRNAs, we could build the bias ratings of disease to miRNA based on the known miRNA–disease association. To achieve this purpose, the agglomerative hierarchical clustering was used to construct the bias ratings by taking advantage of the integrated miRNA similarity.

Specifically, we clustered all the miRNAs associated with disease d_i based on the agglomerative hierarchical clustering by adopting a bottom-up strategy which deemed each miRNA as a single cluster at the beginning of the process, then merged the other clusters based on the linkage criterion of Ward's minimum variance method (Joe and Ward, 1963). We could denote the distance $RD(m_i, m_j)$ between two miRNAs and the distance $DD(d_i, d_j)$ between two diseases which would be used in the linkage criterion with their similarities as follows:

$$RD(m_i, m_j) = 1 - SM(m_i, m_j) \tag{13}$$

$$DD(d_i, d_j) = 1 - SD(d_i, d_j) \tag{14}$$

After clustering, we cut the hierarchical clustering tree to obtain appropriate clusters with an optimal threshold [experimentally set value as 1.1 according to previous work of Shi *et al.* (2015)] which was a distances criterion between clusters. In experiments, we can take advantage of the hierarchical clustering computation package of programming language such as python. Naturally, we denoted the bias ratings from the given disease d_i to the related miRNA m_j as follows:

$$r(d_i, m_j) = n_{cr}/T(d_i) \tag{15}$$

where n_{cr} was the number of the miRNAs in the cluster cr which contained the miRNA m_j. $T(d_i)$ was the total number of the miRNAs associated with disease d_i.

The baseline bipartite network projection recommendation algorithm (Zhou *et al.*, 2007) is a two-round resource transfer process as shown in the second step of Figure 1 which always do not consider bias rating by just marking the disease-related miRNA with value 1, otherwise 0. Therefore, we introduced the bias ratings constructed above to the baseline algorithm. However, different diseases have different bias rating range, which will lead to inconsistent transfer weights in the resource allocation process. To overcome this shortage, we first normalized the original bias rating $r(d_i, m_j)$ with the average of the ratings related with the given disease d_i as follows:

$$\hat{r}(d_i, m_j) = \frac{r(d_i, m_j)}{\bar{r}(d_i)} \tag{16}$$

where

$$\bar{r}(d_i) = \frac{\sum_{j=1}^{nm} r(d_i, m_j)}{T(d_i)} \tag{17}$$

According to the normalized bias ratings, we introduced the following process which recommends all the potential miRNAs to the given disease with the resource scores. Then we could recommend miRNAs for other diseases in the same way as the process above. In reverse, we could recommend diseases for miRNAs according to similar rules.

Specifically, we firstly allocated the initial resource to the miRNA m_j associated with the given disease d_i as follows:

$$R_{ini}(m_j) = \hat{r}(d_i, m_j) \tag{18}$$

This initial resource allocation tried to emphasize the distinction of disease bias to different miRNAs. Thus, the initial resource allocation became more distinguishable and accordant with disease bias ratings. In the following, every miRNA would be allocated with a resource allocation score after the two-round resource distribution which represented the recommendation power of the miRNA to the

Downloaded from https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/bty333/4986418
by Xiamen University user
on 07 June 2018

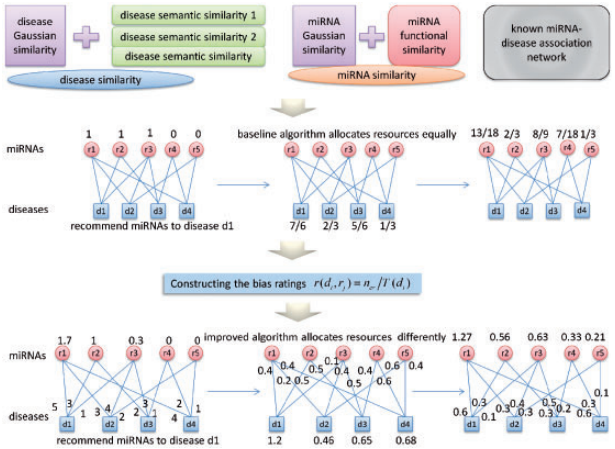


Fig. 1. The flow diagram of BNPMDA model. The first step constructs the known miRNA–disease association network, the disease similarity network and the miRNA similarity network. The second step introduces the baseline bipartite recommendation algorithm by taking the process of recommending miRNAs to disease d_1 as an example. In the third step, the bias ratings are constructed based on the three networks in the first step. Finally, the improved bipartite recommendation algorithm is implemented based on the constructed bias ratings

given disease d_i (see Fig. 1). In the first round, we assigned a transfer weight to each of the resource allocation links from miRNAs to diseases. Specifically, the initial resource transferred from miRNA m_j to disease d_i was calculated as follows:

$$R(d_i, m_j) = \frac{\widehat{r}(d_i, m_j)}{\sum_{k=1}^{nd} \widehat{r}(d_k, m_j)} \times R_{ini}(m_j) \tag{19}$$

Then the disease d_i obtained the allocated resource by adding the contributions from all miRNAs associated to it as follows:

$$R(u_i) = \sum_{j=1}^{nm} R(d_i, m_j) \tag{20}$$

In the second round, we allocated the resource of diseases obtained in the first round back to the miRNAs according to transfer weights from diseases to miRNAs. The transfer weight from disease d_i to miRNA m_j equaled to the ratio of the disease rating $r(d_i, m_j)$ to the sum of ratings from the disease d_i to all the miRNAs. Different from the first round, no normalization needed to be done in this round because all these transfer weights come from the bias ratings of the same disease. Hence, we could formulate the final resource allocation to the miRNA m_j by

$$R_{fin}(v_j) = \sum_{i=1}^{nd} R_{fin}(d_i, m_j) \tag{21}$$

where $R_{fin}(d_i, m_j)$ denoted the final resource allocated from disease d_i to miRNA m_j, which was calculated as follows:

$$R_{fin}(d_i, m_j) = \frac{r(d_i, m_j)}{\sum_{j=1}^{nm} r(d_i, m_j)} \times R(d_i) \tag{22}$$

According to this formula, the rating-based resource distribution allocated the disease resource discriminatively. This process considered the particularity of every transfer link so that it could improve the accuracy of miRNA–disease association prediction.

Similar to the disease-oriented bipartite recommendation which recommended miRNAs to diseases with the resource allocation scores, we in reverse implemented the miRNA-oriented bipartite recommendation which recommended diseases to miRNAs and obtained the final resource from miRNA m_j to disease d_i which was denoted as $R_{fin}(m_j, d_i)$. Furthermore, we combined the disease-oriented final resource $R_{fin}(d_i, m_j)$ and the miRNA-oriented final resource $R_{fin}(m_j, d_i)$ to generate the final prediction score of association between disease d_i and miRNA m_j as follows:

$$FPS(d_i, m_j) = \frac{R_{fin}(d_i, m_j) + R_{fin}(m_j, d_i)}{2} \tag{23}$$

3 Results

3.1 Performance evaluation

Three types of cross validation were implemented to evaluate the prediction accuracy of BNPMDA, which included Global LOOCV, Local LOOCV and 5-fold cross validation (see Fig. 2). In Global LOOCV, each known miRNA–disease association was left out in turn to be considered as test sample and the other remaining known associations were considered as training samples. Then we carried out the BNPMDA to predict the scores of all the disease-miRNA associations and compared the score of the test association with the scores of the other candidate associations to observe whether its ranking was greater than the given threshold. Different from the Global LOOCV, the Local LOOCV only considered the ranking of the score generated by the test association among the candidate associations which were merely related to the investigated disease. According to the results of the Global and Local LOOCV, ROC curve was drew to present the visible accuracy description by plotting true positive rate (TPR, sensitivity) versus false positive rate (FPR, 1-specificity) at different thresholds. Specifically, sensitivity refers to the percentage of the true positive samples whose ranking is higher than the given threshold in the whole positive samples. Meanwhile, specificity denotes the percentage of negative samples with rankings lower than the given threshold in the whole negative samples. AUC was further calculated to demonstrate the prediction ability of BNPMDA. If the AUC equals to 1, it indicates that the model has perfect prediction performance. If the AUC equals to 0.5, it indicates that the model only has random prediction performance. As a result, BNPMDA obtained the AUC of 0.9028 in the Global LOOCV, and the AUC of 0.8380 in the Local LOOCV as shown in Figure 2. To compare the performance of our method with the previous methods, we implemented the methods of HGIMDA, RLSMDA, HDMP, WBSMDA and RWRMDA on the same dataset. As a result, the AUCs of HGIMDA, RLSMDA, HDMP and WBSMDA in Global LOOCV were 0.8781, 0.8426, 0.8366 and 0.8030, respectively. In Local LOOCV, their AUCs were 0.8077, 0.6953, 0.7702 and 0.8031, respectively. Differently, RWRMDA only had AUC of local LOOCV (0.7891) which was one of its defects because it could not simultaneously uncover the missing associations for all diseases.

In the more rigorous manner, 5-fold cross validation was also implemented to further estimate the prediction accuracy of the BNPMDA model by randomly dividing the whole known associations equally into five parts and treating each one of the five parts as test samples in turn by removing the associations of the current test samples simultaneously. Afterwards, every test sample would be scored and compared with scores of the candidate miRNA–disease pairs to obtain the rankings among them. We repeated this procedure 100 times to obtain a more accurate average AUC value. By comparing with the previous models of RLSMDA, HDMP and

Downloaded from https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/bty333/4986418
by Xiamen University user
on 07 June 2018

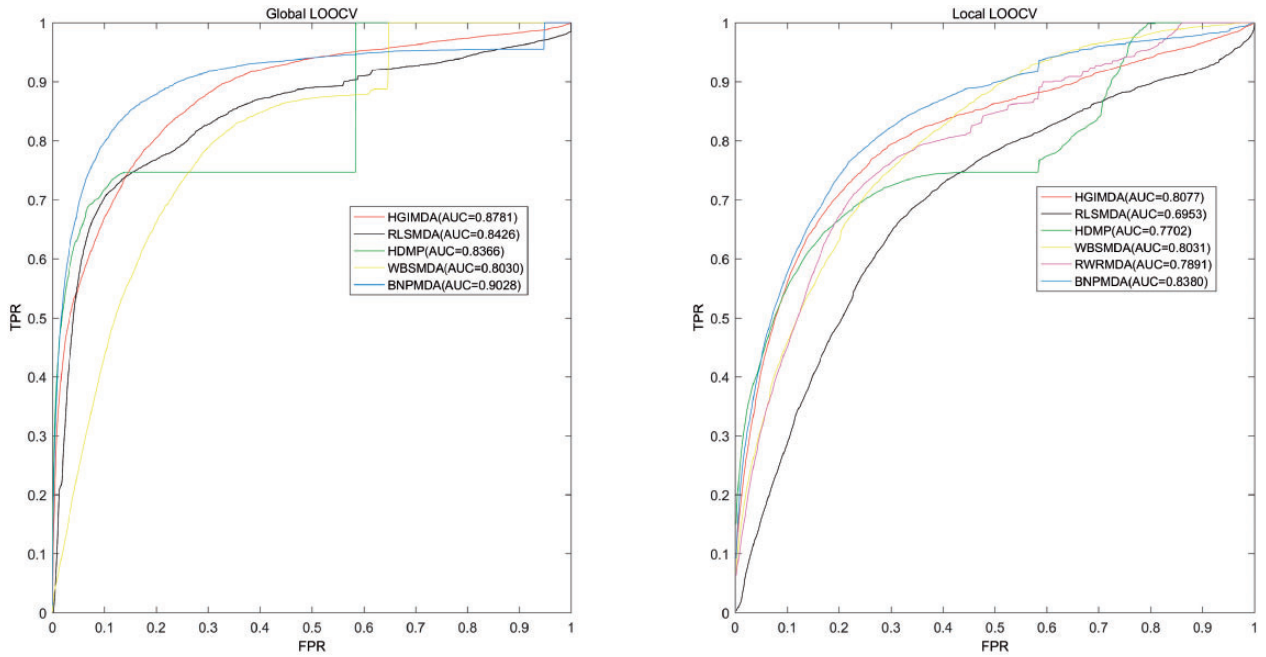


Fig. 2. AUC of global LOOCV (left) compared with HGIMDA, RLSMDA, HDMP and WBSMDA; AUC of local LOOCV (right) compared with HGIMDA, RLSMDA, HDMP, WBSMDA and RWRMDA

WBSMDA, whose average AUCs were 0.8569 ± 0.0020, 0.8342 ± 0.0010 and 0.8185 ± 0.0009 respectively, we could further confirm the effectiveness and the high accuracy of BNPMDA with the average AUC of 0.8980 ± 0.0013 for potential miRNA–disease associations prediction.

3.2 Case studies

We studied three major popular diseases of human beings based on another two miRNA–disease databases, namely dbDEMC database and miR2Disease database. We observed the number of miRNAs that were verified at least one of the three diseases respectively in the top 10, top 20 and even top 50 in the two databases.

Colon neoplasm is the second most common cause of death from cancer. Meanwhile, it is also one of the best-understood neoplasms from a genetic perspective and a class of diseases characterized by out-of-control cell growth. Colon neoplasm forms when this uncontrolled cell growth happens in the cells of the large intestine (Manfredi, 2014). Most colon neoplasms due to old age and lifestyle factors with only a small number of cases due to underlying genetic disorders (Schwartz, 1975). Increasing numbers of evidences have shown that several miRNAs are highly correlated with colon neoplasm and play important roles in the development of colon neoplasm. For example, the study of Espinosa *et al.* has implicated that several miRNAs in tumorigenesis such as miR-143 and miR-145 have been shown to be constantly down-regulated in colon neoplasm (Stahlhut Espinosa and Slack, 2006), and recent studies of Zhang *et al.* have shown that miR-21, miR-17 and miR-19a induced by phosphatase of regenerating liver-3 promote the proliferation and metastasis of colon neoplasm (Zhang *et al.*, 2012). Through implementing BNPMDA, we obtained the total rankings of the candidate miRNAs. As the result shown, among the top 10, 20 and 50 potential colon neoplasm-related miRNAs, there were 9, 19 and 45 miRNA–disease associations proved by recent experiments reports, respectively (see Table 1). For a further observation, the hsa-mir-21 which ranked first in our prediction results had been proved that it induced stemness by down-regulating transforming growth factor beta receptor 2 (TGFbetaR2) in colon neoplasm cells (Yu *et al.*, 2012).

Esophageal neoplasm is a popular cancer with the rapidly diet transformation of people nowadays. But it is a deadliest cancer which is rarely studied worldwide. Fortunately, recent advances in the diagnosis, staging and treatment of this neoplastic condition have led to small but significant improvements in survival (Enzinger and Mayer, 2003). According to the clinical manifestation, the statistical results have shown that the esophageal neoplasm is age-specific which means that the incidence and mortality rates increased with age (Zeng *et al.*, 2016). Recent researches have shown that the expression of miRNAs has tight associations with the development of esophageal neoplasm. For example, the study of Liu *et al.* showed that the expressions of miR-155, miR-183 and miR-20a in esophageal tissue were found to be significantly associated with increased risk for esophageal neoplasm. Circulating miR-155 was found to have significant diagnostic value for esophageal neoplasm as evidenced by an AUC of 66% (Liu *et al.*, 2012). In view of the aforementioned facts, we implemented the BNPMDA to identify potential related miRNAs for esophageal neoplasms based on known associations in the HMDD database. As a result, 9 out of the top 10 and 45 out of the top 50 predicted miRNAs related to esophageal neoplasms were experimentally confirmed by reports from dbDEMC (see Supplementary Table S1). Among the prediction results of top 50, the first was hsa-mir-17 which had been shown that it closely correlated with the occurrence and progression of esophageal squamous carcinoma and might be used as an indicator for esophageal squamous carcinoma prognosis according to the study of Guo *et al.* (2008).

Lymphoma is a sort of lymph cancer which is developed from blood cell tumors of lymphocytes. Some of the cells in the lymphatic system grow abnormally and out of control when lymphoma occurs. People will have symptom of enlarged lymph nodes, fever, drenching sweats and so on when they suffer from the lymphoma. Eventually,

Downloaded from https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/bty333/4986418
by Xiamen University user
on 07 June 2018

Table 1. Prediction of the top 50 miRNAs associated with colon neoplasms

miRNA	Evidence	miRNA	Evidence
hsa-mir-21	dbdemc; miR2Disease	hsa-mir-150	Unconfirmed
hsa-mir-155	dbdemc; miR2Disease	hsa-mir-199a	Unconfirmed
hsa-mir-20a	dbdemc; miR2Disease	hsa-mir-146b	Unconfirmed
hsa-mir-19b	dbdemc; miR2Disease	hsa-mir-181a	dbdemc; miR2Disease
hsa-mir-18a	dbdemc; miR2Disease	hsa-mir-29c	Dbdemc
hsa-mir-143	dbdemc; miR2Disease	hsa-mir-183	dbdemc; miR2Disease
hsa-mir-34a	dbdemc; miR2Disease	hsa-mir-182	dbdemc; miR2Disease
hsa-mir-92a	unconfirmed	hsa-mir-93	dbdemc; miR2Disease
hsa-mir-19a	dbdemc; miR2Disease	hsa-mir-200b	Dbdemc
hsa-mir-146a	dbdemc	hsa-mir-223	dbdemc; miR2Disease
hsa-mir-125b	dbdemc	hsa-mir-141	dbdemc; miR2Disease
hsa-mir-29b	dbdemc; miR2Disease	hsa-mir-181b	dbdemc; miR2Disease
hsa-mir-29a	dbdemc; miR2Disease	hsa-let-7f	dbdemc; miR2Disease
hsa-let-7a	dbdemc; miR2Disease	hsa-mir-196a	dbdemc; miR2Disease
hsa-mir-16	dbdemc	hsa-mir-205	Dbdemc
hsa-mir-106b	dbdemc; miR2Disease	hsa-mir-34c	miR2Disease
hsa-mir-222	dbdemc	hsa-let-7b	dbdemc; miR2Disease
hsa-mir-31	dbdemc; miR2Disease	hsa-let-7c	Dbdemc
hsa-mir-221	dbdemc; miR2Disease	hsa-let-7d	Dbdemc
hsa-mir-9	dbdemc; miR2Disease	hsa-mir-200a	Unconfirmed
hsa-mir-133a	dbdemc; miR2Disease	hsa-mir-199b	Dbdemc
hsa-mir-15a	dbdemc	hsa-mir-107	dbdemc; miR2Disease
hsa-mir-214	dbdemc	hsa-mir-125a	dbdemc; miR2Disease
hsa-mir-200c	dbdemc; miR2Disease	hsa-mir-7	dbdemc; miR2Disease
hsa-mir-1	dbdemc; miR2Disease	hsa-mir-210	Dbdemc

Note: The first column records top 1–25 related miRNAs. The second column records the top 26–50 related miRNAs.

the cancerous cells may form a tumor which continues to grow as the cancerous cells reproduce (Alizadeh *et al.*, 2000). Recent experimental studies showed that the expression level of miR-138, miR-29c, miR26a and miR-16 was found to be reduced in t(14; 18)–negative follicular lymphoma (FL). Particularly, the down-regulation of miR-16 in t(14; 18)–negative FL was statistically significant (Leich *et al.*, 2011). Therefore, it is necessary and important to take lymphoma as a case for the additional prediction. After the implementation of BNPMDA, 9 out of top 10 potential lymphoma-associated miRNAs in the prediction list have been verified by the recent studies (see Supplementary Table S2). Furthermore, for the top 50 verified lymphoma-associated miRNAs predicted by BNPMDA, 44 of them have experimental literature evidences. For example, the study of Akao *et al.* had confirmed the down-regulation of miR-143 and miR-145 in B-cell malignancies (Akao *et al.*, 2007).

For demonstrating all the results of prediction, we showed the prediction scores of all the candidate miRNA–disease pairs obtained from BNPMDA model. This table contains the potential miRNAs associated with all the human diseases in HMDD database (see Supplementary Table S3).

We also tested our model on the old version of the HMDD to see whether the BNPMDA still performed well on it. Through the experiment, the effectiveness of BNPMDA on predicting potential miRNA–disease associations had been confirmed based on three different databases including dbDEMC, miR2Disease and the latest version of HMDD. For instance, there were 10, 20 and 48 miRNAs out of top 10, 20 and 50 miRNAs related with breast neoplasms confirmed by three databases mentioned above (see Table 2). By taking the hsa-mir-16 as an example which ranked the first in the top 50, research of Hu *et al.* had shown that Serum hsa-mir-16 was consistently differentially expressed between breast cancer cases and controls (Hu *et al.*, 2012).

4 Discussion and conclusion

In this paper, we proposed the Bipartite Network Projection for MiRNA–Disease Association prediction (BNPMDA) based on known miRNA–disease associations, miRNA functional similarity, disease semantic similarity and Gaussian interaction profile kernel similarity. We took advantage of the agglomerative hierarchical clustering to construct the bias rating from disease to miRNA and the bias rating from miRNA to disease based on the integrated miRNA similarity and the integrated disease similarity at the beginning of the model. Furthermore, we improved the baseline algorithm of bipartite network recommendation based on the constructed bias ratings by reflecting the distinctness of preference in the resource allocation process from a given disease to various miRNAs or from a given miRNA to various diseases. Furthermore, three types of cross validation and several case studies on important human diseases have been implemented. As a result, BNPMDA performed well both in the cross validations and the case studies.

The excellent performance of BNPMDA mainly attributes to the following several important factors. Firstly, the baseline algorithm of bipartite recommendation outperforms some previous classical methods to recommend the items to users in the field of business. This superiority guaranteed the basic effectiveness of our proposed method. Secondly, the improved bipartite recommendation algorithm could reflect the bias ratings of the diseases and the miRNAs to the resource initialization process and the resource transfer process. This kind of rating-integrated algorithm could take full advantage of the similarity information of both the miRNAs and the diseases to further improve the accuracy of the prediction. Last but not least, increasing numbers of disease-miRNA association data have been discovered these years due to the rapid development of the biological experiment technology.

Downloaded from https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/bty333/4986418
by Xiamen University user
on 07 June 2018

Table 2. Prediction of the top 50 miRNAs associated with breast neoplasms

miRNA	Evidence	miRNA	Evidence
hsa-mir-16	dbdemc; HMDD	hsa-mir-15b	Dbdemc
hsa-let-7e	dbdemc; HMDD	hsa-mir-203	dbdemc; miR2Disease; HMDD
hsa-let-7b	dbdemc; HMDD	hsa-mir-130a	Dbdemc
hsa-mir-223	dbdemc; HMDD	hsa-mir-30e	Unconfirmed
hsa-mir-92a	HMDD	hsa-mir-18b	dbdemc; HMDD
hsa-let-7i	dbdemc; miR2Disease; HMDD	hsa-mir-23b	dbdemc; HMDD
hsa-let-7c	dbdemc; HMDD	hsa-mir-100	dbdemc; HMDD
hsa-let-7g	dbdemc; HMDD	hsa-mir-142	Unconfirmed
hsa-mir-126	dbdemc; miR2Disease; HMDD	hsa-mir-196b	Dbdemc
hsa-mir-106a	dbdemc	hsa-mir-224	dbdemc; HMDD
hsa-mir-181a	dbdemc; miR2Disease; HMDD	hsa-mir-192	Dbdemc
hsa-mir-92b	dbdemc	hsa-mir-198	Dbdemc
hsa-mir-24	dbdemc; HMDD	hsa-mir-135a	dbdemc; HMDD
hsa-mir-150	dbdemc	hsa-mir-22	dbdemc; miR2Disease; HMDD
hsa-mir-195	dbdemc; miR2Disease; HMDD	hsa-mir-372	Dbdemc
hsa-mir-101	dbdemc; miR2Disease; HMDD	hsa-mir-98	dbdemc; miR2Disease
hsa-mir-191	dbdemc; miR2Disease; HMDD	hsa-mir-31	dbdemc; miR2Disease; HMDD
hsa-mir-107	dbdemc; HMDD	hsa-mir-124	dbdemc; HMDD
hsa-mir-99b	dbdemc	hsa-mir-424	Dbdemc
hsa-mir-182	dbdemc; miR2Disease; HMDD	hsa-mir-335	dbdemc; miR2Disease; HMDD
hsa-mir-29c	dbdemc; miR2Disease; HMDD	hsa-mir-95	Dbdemc
hsa-mir-30a	miR2Disease; HMDD	hsa-mir-212	Dbdemc
hsa-mir-199b	dbdemc; HMDD	hsa-mir-27a	dbdemc; miR2Disease; HMDD
hsa-mir-373	dbdemc; miR2Disease; HMDD	hsa-mir-128b	miR2Disease
hsa-mir-32	dbdemc	hsa-mir-181d	dbdemc; miR2Disease

Note: The first column records top 1–25 related miRNAs. The second column records the top 26–50 related miRNAs.

However, there are still some limitations in this model. First of all, though known miRNA–disease association data have been more than before, it is still a small quantity for the prediction to obtain enough accurate results. Secondly, because the bipartite recommendation algorithm is implemented by allocating the resource based on the known associations between different miRNAs and diseases, the BNPMDA is not suitable for the prediction of diseases without any known associated miRNAs. This shortage limits the application range of the BNPMDA.

Acknowledgements

We thank anonymous reviewers for very valuable suggestions.

Funding

X.C. was supported by National Natural Science Foundation of China under Grant Nos. 61772531 and 11631014. Q.Z. was supported by the Doctor Startup Foundation from Liaoning province under Grant No. 20170520217 and Innovation Team Project from the Education Department of Liaoning Province under Grant No. LT2015011. H.L. was supported by National Natural Science Foundation of China under Grant No. 31570160, Important Scientific and Technical Achievements Transformation Project under Grant Z17-5-078, Large-scale Equipment Shared Services Project under Grant No. F15165400 and Applied Basic Research Project under Grant No. F16205151.

Conflict of Interest: none declared.

References

Akao,Y. *et al*. (2007) Downregulation of microRNAs-143 and -145 in B-cell malignancies. *Cancer Sci*., **98**, 1914–1920.

Alizadeh,A.A. *et al*. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–511.

Bandyopadhyay,S. *et al*. (2010) Development of the human cancer microRNA network. *Silence*, **1**, 6.

Bartel,D.P. (2009) MicroRNAs: target recognition and regulatory functions. *Cell*, **136**, 215–233.

Chen,X. *et al*. (2012) RWRMDA: predicting novel human microRNA–disease associations. *Mol. Biosyst*., **8**, 2792–2798.

Chen,X. *et al*. (2015) RBMMMDA: predicting multiple types of disease-microRNA associations. *Sci. Rep*., **5**, 13877.

Chen,X. and Yan,G.Y. (2015) Semi-supervised learning for potential human microRNA–disease associations inference. *Sci. Rep*., **4**, 5501–5501.

Chen,X. *et al*. (2016a) HGIMDA: heterogeneous graph inference for miRNA–disease association prediction. *Oncotarget*, **7**, 65257–65269.

Chen,X. *et al*. (2016b) A novel approach based on KATZ measure to predict associations of human microbiota with non-infectious diseases. *Bioinformatics*, **7**, 733–739.

Chen,X. *et al*. (2016c) WBSMDA: within and between score for MiRNA–Disease Association prediction. *Sci. Rep*., **6**, 21106.

Enzinger,P.C. and Mayer,R.J. (2003) Esophageal cancer. *N. Engl. J. Med*., **349**, 2241–2252.

Gokhale,A. *et al*. (2010) Distinctive microRNA signature of medulloblastomas associated with the WNT signaling pathway. *J. Cancer Res. Therap*., **6**, 521–529.

Guo,Y. *et al*. (2008) Distinctive microRNA profiles relating to patient survival in esophageal squamous cell carcinoma. *Cancer Res*., **68**, 26–33.

Hu,Z. *et al*. (2012) Serum microRNA profiling and breast cancer risk: the use of miR-484/191 as endogenous controls. *Carcinogenesis*, **33**, 828–834.

Jiang,Q. *et al*. (2010) Prioritization of disease microRNAs through a human phenome-microRNAome network. *BMC Syst. Biol*., **4**, S2.

Joe,H. and Ward,J. (1963) Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc*., **58**, 236–244.

Jopling,C.L. *et al*. (2005) Modulation of hepatitis C virus RNA abundance by a liver-specific MicroRNA. *Science*, **309**, 1577–1581.

Kalinowski,F.C. *et al*. (2012) Regulation of epidermal growth factor receptor signaling and erlotinib sensitivity in head and neck cancer cells by miR-7. *PloS One*, **7**, e47067–e47576.

Lee,R.C. *et al*. (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, **89**, 1828–1835.

Downloaded from https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/bty333/4986418
by Xiamen University user
on 07 June 2018

Leich,E. *et al*. (2011) MicroRNA profiles of t(14; 18)–negative follicular lymphoma support a late germinal center B-cell phenotype. *Blood*, **118**, 5550–5558.

Li,X. *et al*. (2011) Prioritizing human cancer microRNAs based on genes' functional consistency between microRNA and cancer. *Nucleic Acids Res*., **39**, e153–e153.

Li,Y. *et al*. (2014) HMDD v2.0: a database for experimentally supported human microRNA and disease associations. *Nucleic Acids Res*., **42**, D1070–D1074.

Liu,R. *et al*. (2012) Circulating miR-155 expression in plasma: a potential biomarker for early diagnosis of esophageal cancer in humans. *J. Toxicol. Environ. Health A*, **75**, 1154–1162.

Lu,M. *et al*. (2008) An analysis of human MicroRNA and disease associations. *PloS One*, **3**, e3420.

Mørk,S. *et al*. (2014) Protein-driven inference of miRNA–disease associations. *Bioinformatics*, **30**, 392–397.

Manfredi,S. (2014) Colon cancer: from mass screening to personalised treatment. *Oncologie*, **16**, S484–S484.

Meola,N. *et al*. (2009) microRNAs and genetic diseases. *PathoGenetics*, **2**, 7.

Pasquinelli,A.E. (2016) A sense-able microRNA. *Genes Dev*., **30**, 2019–2020.

Ribeiro,A.O. *et al*. (2014) MicroRNAs: modulators of cell identity, and their applications in tissue engineering. *Microrna*, **3**, 45–53.

Schwartz,M.K. (1975) Enzymes in colon cancer. General information. *Cancer*, **36**, 2334–2336.

Shi,H. *et al*. (2013) Walking the interactome to identify human miRNA–disease associations through the functional link between miRNA targets and disease genes. *BMC Syst. Biol*., **7**, 101.

Shi,J.Y. *et al*. (2015) Predicting drug-target interaction for new drugs using enhanced similarity measures and super-target clustering. *Methods*, **83**, 98–104.

Stahlhut Espinosa,C.E. and Slack,F.J. (2006) The role of microRNAs in cancer. *Yale J. Biol. Med*., **79**, 131–140.

Von,B.M. *et al*. (2012) MicroRNA 15a, inversely correlated to PKCα, is a potential marker to differentiate between benign and malignant renal tumors in biopsy and urine samples. *Am. J. Pathol*., **180**, 1787–1797.

Wang,D. *et al*. (2010) Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. *Bioinformatics*, **26**, 1644–1650.

Wightman,B. *et al*. (1993) Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in *C. elegans*. *Cell*, **75**, 855–862.

Wong,L. *et al*. (2015) Detection of interactions between proteins through rotation forest and local phase quantization descriptors. *Int. J. Mol. Sci*., **17**, 21.

Xu,C. *et al*. (2014) Prioritizing candidate disease miRNAs by integrating phenotype associations of multiple diseases with matched miRNA and mRNA expression profiles. *Mol. Biosyst*., **10**, 2800–2809.

Xu,J. *et al*. (2011) Prioritizing candidate disease miRNAs by topological features in the miRNA target-dysregulated network: case study of prostate cancer. *Mol. Cancer Ther*., **10**, 1857–1866.

Xuan,P. *et al*. (2013) Prediction of microRNAs associated with human diseases based on weighted k most similar neighbors. *PLoS One*, **8**, e70204.

Yu,Y. *et al*. (2012) MicroRNA-21 induces stemness by downregulating transforming growth factor beta receptor 2 (TGFβR2) in colon cancer cells. *Carcinogenesis*, **33**, 68–76.

Zeng,H.M. *et al*. (2016) Esophageal cancer statistics in China, 2011: estimates based on 177 cancer registries. *Thorac. Cancer*, **7**, 232–237.

Zhang,J. *et al*. (2012) miR-21, miR-17 and miR-19a induced by phosphatase of regenerating liver-3 promote the proliferation and metastasis of colon cancer. *Br. J. Cancer*, **107**, 352–359.

Zhou,T. *et al*. (2007) Bipartite network projection and personal recommendation. *Phys. Rev. E Stat. Nonlinear Soft Matter Phys*., **76**, 046115.

Downloaded from https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/bty333/4986418
by Xiamen University user
on 07 June 2018