Gao et al. BMC Bioinformatics (2020) 21:61
https://doi.org/10.1186/s12859-020-3409-x

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Graph regularized $L_{2,1}$-nonnegative matrix factorization for miRNA-disease association prediction

Zhen Gao, Yu-Tian Wang, Qing-Wen Wu, Jian-Cheng Ni* and Chun-Hou Zheng*

Abstract

Background: The aberrant expression of microRNAs is closely connected to the occurrence and development of a great deal of human diseases. To study human diseases, numerous effective computational models that are valuable and meaningful have been presented by researchers.

Results: Here, we present a computational framework based on graph Laplacian regularized $L_{2,1}$-nonnegative matrix factorization ($GRL_{2,1}$-NMF) for inferring possible human disease-connected miRNAs. First, manually validated disease-connected microRNAs were integrated, and microRNA functional similarity information along with two kinds of disease semantic similarities were calculated. Next, we measured Gaussian interaction profile (GIP) kernel similarities for both diseases and microRNAs. Then, we adopted a preprocessing step, namely, weighted K nearest known neighbours (WKNKN), to decrease the sparsity of the miRNA-disease association matrix network. Finally, the $GRL_{2,1}$-NMF framework was used to predict links between microRNAs and diseases.

Conclusions: The new method ($GRL_{2,1}$-NMF) achieved AUC values of 0.9280 and 0.9276 in global leave-one-out cross validation (global LOOCV) and five-fold cross validation (5-CV), respectively, showing that $GRL_{2,1}$-NMF can powerfully discover potential disease-related miRNAs, even if there is no known associated disease.

Keywords: miRNA, Disease, miRNA-disease associations, NMF $L_{2,1}$-norm

Background

MicroRNAs (miRNAs), which play crucial roles in the regulation of gene expression after transcription in organisms and vegetation, are 17–24 nt noncoding endogenous RNAs [1–3]. In 1993, Lee et al. [4] identified the first microRNA (miRNA) called lin-4 in *Caenorhabditis elegans*. Thereafter, a large number of miRNAs have been identified from a wide variety of species, such as plants, animals, and viruses [5, 6]. MiRNAs are associated with key biological processes, including development, differentiation, programmed cell death and cell proliferation [7, 8]. Past studies have indicated that abnormal miRNA expression participates in the development process of a variety of human diseases [9–11]. However, inferring microRNA-disease connections through manual experiments is tremendously costly, laborious, prone to failure and time consuming. Thus, the development of computation-based methods to infer disease-connected microRNAs is urgently needed, as they could solve the above problems and greatly facilitate human disease diagnosis and treatment [12–15].

For the past few years, in order to explore the pathogenic mechanism of human disease at the small molecule level and design specific molecular instruments for diagnosis treatment and prevention, considerable efforts have been made to develop computational algorithms for inferring disease-associated microRNAs according to the assumptions that microRNAs have similar functions that are highly likely to be connected with similar diseases, and vice versa. Numerous similarity measurement-based approaches according to heterogeneous biological information have been proposed to identify the interactions between microRNAs and diseases. Jiang et al. [16] inferred disease-related miRNAs by prioritizing the whole human miRNAome connected with disease that we investigated based on miRNA functional similarity information as well as the human

* Correspondence: nijch@163.com; zhengch99@126.com
School of Software, Qufu Normal University, Qufu 273165, China

phenome-microRNAome network. Li et al. [17] proposed a computation-based model to infer the possible disease–related miRNAs via calculations of FCS between the disease-gene and the target-gene, which had verification. There is an assumption that if two various diseases have phenotypic connections, they have similar molecular machinery and similar molecular mechanisms. Xu et al. [18] inferred human disease-connected microRNAs by fusing experimentally verified human disease genes as well as context-dependent miRNA-target interactions to prioritize disease-connected microRNAs. In line with weighted k nearest neighbours, HDMP was proposed by Xuan et al. [19] for identifying potential miRNA-disease associations. They presented a measurement method including the details of the disease term along with phenotypic similarities among diseases for the purpose of measuring the miRNA functional similarities. In addition, considering the miRNAs of the same miRNA family or cluster and their relationship to a group of diseases, they were given a higher weight. However, HDMP is not appropriate for diseases that have sparse connections with miRNAs. Chen et al. [20] developed miRPD in which experimentally verified or predicted interactions between miRNAs and proteins as well as text-extracted connections between protein and disease associations were explicitly utilized to calculate the probability that a microRNA-disease association exists. Chen et al. [21] developed WBSMDA according to the calculation of the within-scores and between-scores of every miRNA-disease group to identify potential disease-related miRNAs. Take a miRNA as an example, there is a miRNA set A whose elements all have known connections with the investigated disease d. The propose of within-scores is finding a miRNA in set A that has the highest similarity score with the investigated miRNA. There is a miRNA set B whose elements all have unknown connections with the investigated disease d. The proposed between-scores involves finding a miRNA in set B that has the highest similarity score with the investigated miRNA. Chen et al. [22] developed HGIMDA through an iteration approach in line with a graph that consists of many different types of bioinformatics information, such as the functional similarities of microRNAs, semantic similarities of diseases, kernel similarity of Gaussian interaction profiles and experimental verification of microRNA-disease connections. Yu et al. [23] proposed an assembled identification approach to infer potential microRNA-disease associations by modifying the existing maximizing information flow methods based on integrated microRNA functional similarity information, disease semantic information and phenotypic similarity information; these potential associations along with manually validated microRNA-disease interactions were placed into a phenome-microRNAome network. Chen et al. [24] presented a novel framework called RKNNMDA that utilizes ranking and k nearest neighbours. They integrated the functional similarity of microRNAs, semantic similarity of diseases, kernel similarity of Gaussian interaction profiles and experimental verification of microRNA-disease association and obtained miRNA's (disease's) k nearest neighbours via the KNN model. Next, they implemented the SVM ranking model to re-rank the above k nearest neighbours and thus obtained the eventual rankings of all possible microRNA-disease associations. In addition, RKNNMDA could also predict possible microRNA-disease connections for human diseases that don't have manually validated associated miRNAs. Chen et al. [25] introduced Jaccard similarity among microRNAs and diseases in the BLHARMDA model to identify potential miRNA-disease interactions and then introduced an improved KNN framework into the bipartite local model method. Chen et al. [26] defined all paths between a given miRNA and disease as prediction scores, based on the assumption that if there are more paths between the miRNA and disease, the two are more likely to be related.

In addition, a host of studies in accordance with random walk with restart have been proposed for identifying potential microRNA-disease connections and finally obtained good predictive behaviour. A random walk with restart was presented by Chen et al. [27], who also integrated the manually verified microRNA-disease association information and functional similarity information of miRNAs. Considering the functional links among microRNA targets and human disease genes in a protein association network, Shi et al. [28] devised a computational model to infer likely microRNA-disease connections. This method utilized global network distance measurement, random walk analysis, and the construction of a microRNA-disease network to investigate microRNA-disease connections from a global perspective. Xuan et al. [29] designed a novel framework named MIDP, which predicted disease-connected miRNAs for diseases with known associated microRNAs in line with random walks. They analysed the attributes of the labelled and unlabelled nodes of the miRNA network and then established transition matrices, whose transition weights between the nodes were proportionate to the similarity between them. Furthermore, they presented an extension method called MIDPE, especially for diseases that don't have manually verified connected microRNAs. Liu et al. [30] proposed a method to identify possible disease-connected microRNAs by utilizing a random walk with restart in accordance with a heterogeneous graph, which was established by combining disease semantic similarities and disease functional similarities, as well as the miRNA similarities that were obtained utilizing microRNA-target gene and microRNA-long noncoding

RNA connections. Luo and Xiao [13] first established a heterogeneous network containing microRNA and disease information and then adopted a bi-random walk model to identify possible microRNA-disease connections. Finally, all microRNA candidates of an investigated disease were ranked.

Furthermore, machine learning-based algorithms, such as support vector machines, have been applied to bioinformatics and computational biology and have improved the prediction performance to some extent [31]. Xu et al. [32] presented MTDN to infer potential microRNA-disease associations. They identified positive disease-related miRNAs from negative samples through the SVM classifier in accordance with the characteristics of microRNA target-dysregulated network topology information. Chen et al. [33] identified miRNA-disease links based on regularized least squares (RLS) for identifying the miRNA-disease links. RLSMDA integrates known disease-microRNA connections, a disease semantic similarities dataset, and a miRNA functional similarities network and is thus suitable for predicting novel miRNAs for diseases without any manually validated connections with microRNAs. Li et al. [34] utilized a matrix completion model in line with manually validated microRNA-disease connections to infer candidates for diseases that did not have any experimentally proven connected microRNAs. In addition, MCMDA does not need negative associations. Chen et al. [35] proposed a random forest-based framework (RFMDA) for microRNA-disease connection prediction. RFMDA identifies possible disease-associated microRNAs by employing the random forest model to identify robust attributes from the miRNA-disease attribute collection. Chen et al. [36] predicted disease-associated miRNAs based on heterogeneous label propagation (HLPMDA), in which heterogeneous data were integrated into a heterogeneous network. Chen et al. [37] inferred disease-associated miRNAs with restricted Boltzmann machine (RBM); this model can acquire both disease-connected miRNAs as well as the corresponding forms of their links. However, this method is not suitable for diseases that do not have any known miRNA-disease associations, and selecting the right parameter values remains a significant issue for RBMMMDA. Chen et al. [38] first integrated a heterogeneous network, then put it into a stacked autoencoder for the purpose of detecting the deep representation of the heterogeneous information, finally utilizing an SVM classifier to prioritize all the candidates. Chen et al. [39] first constructed a feature vector according to the statistics, graph theory and matrix decomposition of the bioinformatics data and then put this vector into EGBMMDA to obtain a regression tree. Chen et al. [40] extracted three kinds of features, namely, statistical features from similarity measurements, graph theoretical features from similarity networks, and matrix factorization results from miRNA-disease associations. Then, disease-related miRNAs were discovered based on a decision tree classifier. Chen et al. [41] predicted disease-connected miRNAs by adopting sparse subspace learning with Laplacian regularization and L_1-norm. Interestingly, they extracted features and constructed objective functions from miRNA and disease perspectives, separately. Chen et al. [42] used a decision tree as a weak classifier and then integrated these weak classifiers into a strong classifier according to weights. It is worth noting that they implemented k-means to balance positive samples and negative samples.

Moreover, many researchers have made promising models with recommendation systems for microRNA-disease connection prediction purposes. Zou et al. [43] proposed two approaches, namely, KATZ and CATAPULT, for identifying miRNA-disease links. In line with the manually verified microRNA-disease link network, microRNA similarities network and disease similarities network, KATZ integrates the social network analysis approach with machine learning. Chen et al. [44] inferred disease-related miRNAs based on ensemble learning and link prediction (ELLPMDA). According to global similarity measures, ELLPMDA uses ensemble learning for integrating ranking results, which were obtained via three typical similarity-measurement approaches. Chen et al. [45] constructed a heterogeneous network and predicted disease-connected miRNAs in line with the rating-integrated bipartite network recommendation as well as experimentally verified miRNA–disease connections.

In addition, a fair number of studies based on matrix factorization have been presented for possible disease-connected microRNA prediction purposes. Zhao et al. [46] presented symmetric nonnegative matrix factorization (SNMFMDA) to infer disease-connected microRNAs with the NMF and Kronecker regularized least square (KronRLS) approaches. Zhong et al. [47] proposed a nonnegative matrix factorization (NMF)-based algorithm to predict disease-related microRNA candidates based on a bilayer network that was constructed with regard to the intricate links among microRNAs, among human diseases and between microRNAs and human diseases. Xiao et al. [48] introduced graph Laplacian regularized into NMF (GRNMF) based on heterogeneous data for inferring potential disease-connected microRNAs, particularly for many diseases without known associations. They introduced a preprocessing step, weighted k nearest neighbour (WKNKN) profiles, for both microRNAs and diseases, into GRNMF. Chen et al. [49] designed an effective algorithm, MDHGI, according to matrix decomposition as well as a heterogeneous graph inference method for inferring potential miRNA-disease connections.

However, these approaches based on matrix factorization ignored the sparsity of the miRNA-disease association matrix Y, so we utilized a pre-processing step named weighted K nearest known neighbours (WKNKN) [50] to convert the value of the miRNA-disease associations matrix Y into a decimal between 0 and 1. In addition, unlike the traditional nonnegative matrix factorization (NMF) methods, we added $L_{2,1}$-norm as well as GIP (Gaussian interaction profile) kernels into the NMF model. The $L_{2,1}$-norm was added to increase the disease matrix sparsity and eliminate unattached disease pairs [51–53]. Moreover, Tikhonov regularization was added to penalize the non-smoothness of W and H [48, 54, 55], and the graph regularization was primarily intended to assure local-based representation by leveraging the geometry of the data [56].

In this study, we present a computational algorithm based on graph regularized $L_{2,1}$-nonnegative matrix factorization ($GRL_{2,1}$-NMF) to infer the possible connections between microRNAs and diseases in heterogeneous omics data. First, we integrated manually validated microRNA-disease connection information, miRNA functional similarity information and two kinds of disease semantic similarity information, and then we calculated the GIP kernel similarities for the diseases and miRNAs. Then, we utilized WKNKN to decrease the sparsity of matrix Y. Furthermore, we added Tikhonov (L_2), graph Laplacian regularization terms and the $L_{2,1}$-norm to the standard NMF model for predicting disease-associated miRNAs. Finally, five-fold cross validation and global leave-one-out cross validation were implemented to evaluate the effectiveness of our model, and we obtained AUCs of 0.9276 and 0.9280, respectively. Furthermore, we performed case studies on three high-risk human diseases (prostate neoplasms, lung neoplasms and breast neoplasms). As a result, 48, 45 and 45 out of the top 50 likely connected miRNAs of prostate neoplasms, lung neoplasms and breast neoplasms, respectively, were confirmed by HMDD [10] and dbDEMC [57]. Based on the experimental results, we can clearly see that $GRL_{2,1}$-NMF is a valuable approach for inferring possible miRNA-disease connections.

Results

Effect of parameters on the performance of $GRL_{2,1}$-NMF

In this work, we measured two disease semantic similarities, miRNA functional similarity and GIP similarities for miRNAs and diseases. These two disease semantic similarities were integrated as Eq. (1), and the final disease similarity and miRNA similarity were measured as Eq. (2) and Eq. (3), respectively. We defined six parameters, namely, α_1, α_2, γ_1, γ_2, θ_1 and θ_2, to balance the items in Eq. (1), Eq. (2) and *Eq.* (3). The values of α_1 and α_2 ranged from 0.1, 0.2, 0.3, ... to 0.9. γ_1, γ_2, θ_1 and θ_2 ranged from 0,0.1,0.2, ... 0.9, to 1. We conducted a series of experiments on the above parameters to acquire the effects of these parameters. The experimental results are shown in Table 1 and Table 2.

$$SD1(d_i, d_j) = \alpha_1 S_1^d(d_i, d_j) + \alpha_2 S_2^d(d_i, d_j) \tag{1}$$

$$SD(d_i, d_j) = \begin{cases} \gamma_1 SD1(d_i, d_j) + \gamma_2 GD(d_i, d_j) & d_i \text{ and } d_j \text{ have semantic similarity} \\ GD(d_i, d_j) & \text{otherwise} \end{cases} \tag{2}$$

$$SM(m_i, m_j) = \begin{cases} \theta_1 S^m(m_i, m_j) + \theta_2 GM(m_i, m_j) & m_i \text{ and } m_j \text{ have functional similarity} \\ GM(m_i, m_j) & \text{otherwise} \end{cases} \tag{3}$$

In Table 1, we can see that regardless of how α_1 and α_2 change, the AUC of 5-CV remains 0.9276. Thus, for convenience, we set $\alpha_1 = \alpha_2 = 0.5$. The experimental results of parameters θ_1 and θ_2 that balanced miRNA functional similarity (S^m) and GIP similarity for miRNAs (GM) are shown in Table 2 (a), and the results of parameters γ_1 and γ_2 that balanced disease semantic similarity ($SD1$) and GIP similarity for diseases (GD) are shown in Table 2 (b). Thus, we set $\theta_1 = 1$, $\theta_2 = 0$, $\gamma_1 = 1$, and $\gamma_2 = 0$.

Performance evaluation

To evaluate our model's ability to predict disease-related miRNAs, we compared it with three state-of-art methods (ICFMDA [58], SACMDA [59] and IMCMDA [60]) by implementing two validation frameworks: global leave-one-out cross validation (global LOOCV) and five-fold cross validation (5-CV) according to the experimentally validated disease-related miRNAs in HMDD v2.0,

Table 1 The effects of parameters α_1 and α_2 on the results of $GRL_{2,1}$-NMF $\gamma_1 = 1, \gamma_2 = 0$, $\theta_1 = 1$, and $\theta_2 = 0$

	α_1	α_2	AUCs of 5-CV
SD12_1	0.1	0.9	0.9276
SD12_2	0.2	0.8	0.9276
SD12_3	0.3	0.7	0.9276
SD12_4	0.4	0.6	0.9276
SD12_5	0.5	0.5	0.9276
SD12_6	0.6	0.4	0.9276
SD12_7	0.7	0.3	0.9276
SD12_8	0.8	0.2	0.9276
SD12_9	0.9	0.1	0.9276

Table 2 The effects of parameters θ_1,θ_2, γ_1, and γ_2 on the results of $GRL_{2,1}$-NMF (a) $\alpha_1 = 0.5$, $\alpha_2 = 0.5$, $\gamma_1 = 1$, and $\gamma_2 = 0$ (b) $\alpha_1 = 0.5$,$\alpha_2 = 0.5$, $\theta_1 = 1$,and $\theta_2 = 0$

	θ_1	θ_2	AUCs of 5-CV		γ_1	γ_2	AUC of 5-CV
SMGM_1	0.1	0.9	0.9263	SDGD_1	0.1	0.9	0.9276
SMGM_2	0.2	0.8	0.9264	SDGD_2	0.2	0.8	0.9276
SMGM_3	0.3	0.7	0.9267	SDGD_3	0.3	0.7	0.9276
SMGM_4	0.4	0.6	0.9268	SDGD_4	0.4	0.6	0.9276
SMGM_5	0.5	0.5	0.9270	SDGD_5	0.5	0.5	0.9276
SMGM_6	0.6	0.4	0.9270	SDGD_6	0.6	0.4	0.9276
SMGM_7	0.7	0.3	0.9271	SDGD_7	0.7	0.3	0.9276
SMGM_8	0.8	0.2	0.9272	SDGD_8	0.8	0.2	0.9276
SMGM_9	0.9	0.1	0.9272	SDGD_9	0.9	0.1	0.9276
SMGM_10	1	0	0.9276	SDGD_10	1	0	0.9276

which gathered plenty of the known miRNA-disease associations [10].

For the global LOOCV, every known miRNA-disease connection was selected in turn for testing, and others that had also been experimentally verified were considered as training sets for the purpose of model training. In addition, all miRNA-disease associations without evidence were regarded as candidate samples. Next, we calculated the prediction score of all associations by implementing $GRL_{2,1}$-NMF and thus obtained the ranking of each test sample compared with that of the candidate samples. We hold our model as efficient if the ranking of each test sample was higher than a certain threshold. We obtained the corresponding true positive rate (TPR, sensitivity) and false positive rate (FPR, 1-specificity) by setting various thresholds. Sensitivity is the proportion of the testing samples whose ranking was higher than the threshold, while 1-specificity calculates the percentage of the testing samples whose ranking was lower than the threshold. Thus, the receiver operating characteristic (ROC) curve can be plotted in line with TPRs and FPRs obtained by different thresholds. Finally, to evaluate the performance and compare it with that of the other models, the areas under the ROC curve (AUCs) were computed. The AUC value is between 0 and 1, and a model whose AUC value is higher has a better performance. The results showed that $GRL_{2,1}$-NMF, ICFMDA, SACMDA and IMCMDA achieved AUC values of 0.9280, 0.9072, 0.8777 and 0.8384, respectively (see Fig. 1). Clearly, $GRL_{2,1}$-NMF obtained the best performance among the four explored methods.

For 5-CV, all known connections between microRNAs and diseases were randomly distributed into five parts, where one part was selected in turn for testing, and the other four parts were used in turn for training. Moreover, all unknown samples were treated as candidate samples. Like the global LOOCV, we finally calculated the ranking of the test sample relative to the candidate set. Considering the possible bias caused by random sample partitioning for performance evaluation, we repeatedly divided the known miRNA disease associations 100 times and obtained the corresponding ROC curves and AUCs in a similar manner to that for LOOCV. The results showed that $GRL_{2,1}$-NMF had the best predictive performance with an average AUC of 0.9276, and ICFMDA, SACMDA

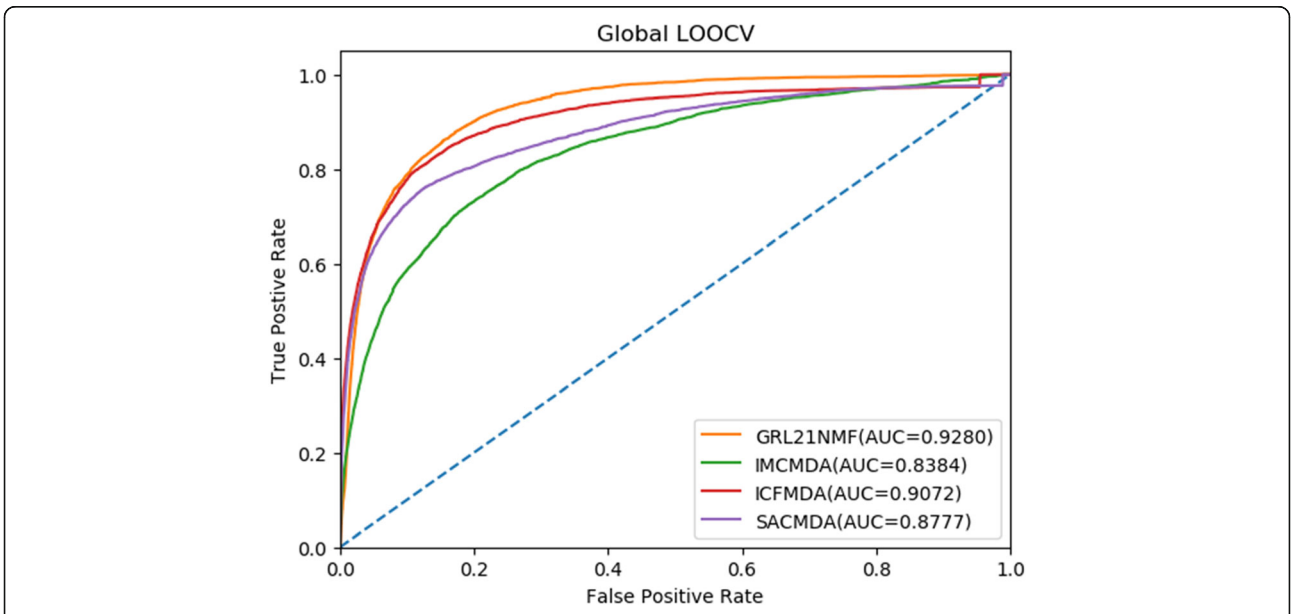

Fig. 1 AUC of global LOOCV compared with those of IMC, ICFMDA and SACMDA

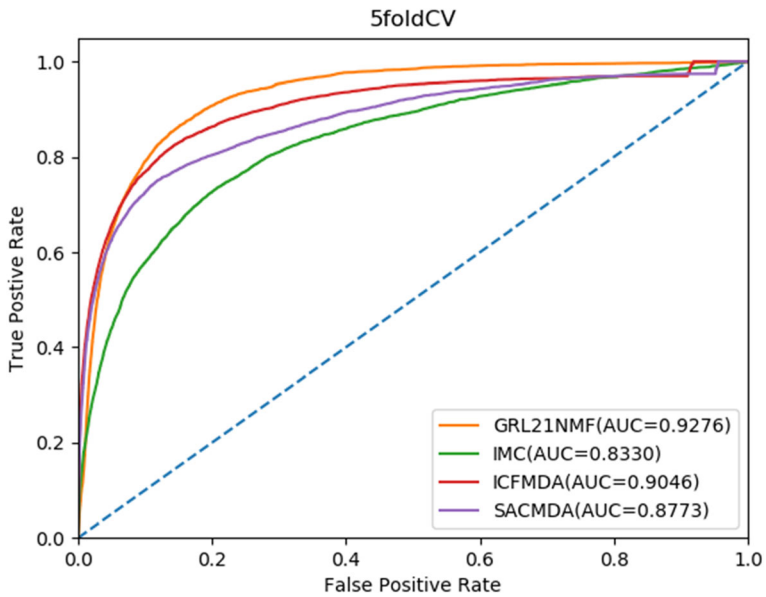

Fig. 2 AUC of 5-fold cross validation compared with those of IMC, ICFMDA and SACMDA

and IMCMDA achieved AUC values of 0.9046, 0.8773 and 0.8330, respectively (see Fig. 2).

Case studies

We constructed a simulation experiment to further demonstrate the effectiveness of $GRL_{2,1}$-NMF for inferring likely disease-connected miRNAs. Here, all manually validated miRNA-disease connections were utilized for prediction, and other associations that did not have evidence were regarded as candidate connections for validation. For every disease, the candidate miRNAs were ranked based on the prediction scores. We used two miRNA-disease databases, namely, HMDD [10] and dbDEMC [57], to verify the inferred possible microRNAs for the investigated disease, including prostate neoplasms, breast neoplasms and lung neoplasms. Finally, the top 50 disease-related miRNAs predicted via $GRL_{2,1}$-NMF are demonstrated in Table 3, Table 4 and Table 5. There are 48,45 and 45 of 50 inferred miRNAs confirmed to have associations with prostate neoplasms, breast neoplasms and lung neoplasms, respectively, by the dbDEMC database and HMDD v3.0 database.

Discussion

Our method, $GRL_{2,1}$-NMF, is an efficient tool for predicting miRNA-disease associations according to the experimental results. The main contributions of this study are listed. First, we added GIP kernel similarities for miRNA and disease associations into the similarity measurement, which improved the dataset reliability. Second, considering the sparsity of observed miRNA-disease associations, we performed a pre-processing step

Table 3 The top 50 potential miRNAs associated with Prostate Neoplasms

miRNA	Evidence	miRNA	Evidence
hsa-mir-1	HMDD; dbDEMC	hsa-mir-32	HMDD; dbDEMC
hsa-mir-21	HMDD; dbDEMC	hsa-let-7i	dbDEMC
hsa-mir-22	HMDD; dbDEMC	hsa-mir-375	HMDD; dbDEMC
hsa-mir-155	HMDD; dbDEMC	hsa-let-7c	HMDD; dbDEMC
hsa-mir-9	HMDD	hsa-mir-200c	HMDD; dbDEMC
hsa-mir-221	HMDD; dbDEMC	hsa-mir-214	HMDD; dbDEMC
hsa-let-7a	dbDEMC	hsa-mir-182	HMDD; dbDEMC
hsa-mir-133a	HMDD; dbDEMC	hsa-mir-106b	HMDD; dbDEMC
hsa-mir-146a	HMDD	hsa-mir-23a	HMDD; dbDEMC
hsa-mir-222	HMDD; dbDEMC	hsa-mir-17	HMDD; dbDEMC
hsa-mir-34a	HMDD; dbDEMC	hsa-let-7e	dbDEMC
hsa-mir-29a	HMDD; dbDEMC	hsa-mir-181	unconfirmed
hsa-mir-142	unconfirmed	hsa-mir-200b	HMDD; dbDEMC
hsa-mir-223	HMDD; dbDEMC	hsa-mir-10b	dbDEMC
hsa-mir-126	HMDD; dbDEMC	hsa-mir-200a	HMDD; dbDEMC
hsa-mir-31	HMDD; dbDEMC	hsa-mir-34c	HMDD
hsa-mir-146b	HMDD; dbDEMC	hsa-mir-205	HMDD; dbDEMC
hsa-mir-29b	HMDD; dbDEMC	hsa-let-7d	HMDD; dbDEMC
hsa-mir-200	HMDD	hsa-mir-210	HMDD; dbDEMC
hsa-mir-143	HMDD; dbDEMC	hsa-mir-192	HMDD; dbDEMC
hsa-mir-16	HMDD; dbDEMC	hsa-mir-196a	HMDD; dbDEMC
hsa-mir-20a	HMDD; dbDEMC	hsa-mir-195	HMDD; dbDEMC
hsa-mir-30a	HMDD	hsa-let-7f	dbDEMC
hsa-let-7b	HMDD; dbDEMC	hsa-mir-181b	HMDD; dbDEMC
hsa-mir-199a	HMDD; dbDEMC	hsa-mir-34b	HMDD

Table 4 The top 50 potential miRNAs associated with Lung Neoplasms

miRNA	Evidence	miRNA	Evidence
hsa-mir-1	HMDD	hsa-mir-139	HMDD; dbDEMC
hsa-mir-181	unconfirmed	hsa-mir-193b	dbDEMC
hsa-mir-200	HMDD	hsa-mir-204	dbDEMC
hsa-mir-26	HMDD	hsa-mir-708	dbDEMC
hsa-mir-195	dbDEMC	hsa-mir-378a	unconfirmed
hsa-mir-92	dbDEMC	hsa-mir-625	dbDEMC
hsa-mir-141	dbDEMC	hsa-mir-367	dbDEMC
hsa-mir-122	HMDD; dbDEMC	hsa-mir-149	HMDD; dbDEMC
hsa-mir-16	HMDD; dbDEMC	hsa-mir-148b	HMDD; dbDEMC
hsa-mir-99a	HMDD; dbDEMC	hsa-mir-328	HMDD; dbDEMC
hsa-mir-129	HMDD; dbDEMC	hsa-mir-302b	dbDEMC
hsa-mir-429	dbDEMC	hsa-mir-302a	dbDEMC
hsa-mir-130a	HMDD; dbDEMC	hsa-mir-373	HMDD; dbDEMC
hsa-mir-451	HMDD; dbDEMC	hsa-mir-92b	dbDEMC
hsa-mir-451a	HMDD; dbDEMC	hsa-mir-23b	dbDEMC
hsa-mir-15b	dbDEMC	hsa-mir-152	HMDD; dbDEMC
hsa-mir-151	unconfirmed	hsa-mir-196b	HMDD; dbDEMC
hsa-mir-15a	HMDD; dbDEMC	hsa-mir-302c	dbDEMC
hsa-mir-151a	unconfirmed	hsa-mir-452	dbDEMC
hsa-mir-296	unconfirmed	hsa-mir-215	HMDD; dbDEMC
hsa-mir-320a	dbDEMC	hsa-mir-302d	dbDEMC
hsa-mir-20b	dbDEMC	hsa-mir-28	dbDEMC
hsa-mir-342	HMDD; dbDEMC	hsa-mir-520a	dbDEMC
hsa-mir-194	HMDD; dbDEMC	hsa-mir-130b	HMDD; dbDEMC
hsa-mir-106b	dbDEMC	hsa-mir-372	HMDD; dbDEMC

Table 5 The top 50 potential miRNAs associated with Breast Neoplasms

miRNA	Evidence	miRNA	Evidence
hsa-mir-1	HMDD; dbDEMC	hsa-mir-330	dbDEMC
hsa-mir-32	HMDD; dbDEMC	hsa-mir-192	HMDD; dbDEMC
hsa-mir-106a	HMDD; dbDEMC	hsa-mir-28	dbDEMC
hsa-mir-26	unconfirmed	hsa-mir-130b	HMDD; dbDEMC
hsa-mir-99a	HMDD; dbDEMC	hsa-mir-211	dbDEMC
hsa-mir-151	HMDD; dbDEMC	hsa-mir-181c	HMDD; dbDEMC
hsa-mir-451	HMDD; dbDEMC	hsa-mir-449a	HMDD; dbDEMC
hsa-mir-92	HMDD; dbDEMC	hsa-mir-449b	dbDEMC
hsa-mir-130a	HMDD; dbDEMC	hsa-mir-99b	dbDEMC
hsa-mir-15b	HMDD; dbDEMC	hsa-mir-208a	HMDD; dbDEMC
hsa-mir-150	HMDD; dbDEMC	hsa-mir-650	dbDEMC
hsa-mir-185	HMDD; dbDEMC	hsa-mir-491	HMDD
hsa-mir-142	HMDD	hsa-mir-532	unconfirmed
hsa-mir-378a	HMDD	hsa-mir-144	HMDD; dbDEMC
hsa-mir-186	dbDEMC	hsa-mir-181d	dbDEMC
hsa-mir-95	dbDEMC	hsa-mir-494	HMDD; dbDEMC
hsa-mir-92b	HMDD; dbDEMC	hsa-mir-362	unconfirmed
hsa-mir-196b	HMDD; dbDEMC	hsa-mir-517a	dbDEMC
hsa-mir-98	HMDD; dbDEMC	hsa-mir-371	dbDEMC
hsa-mir-372	dbDEMC	hsa-mir-371a	unconfirmed
hsa-mir-574	HMDD	hsa-mir-381	HMDD; dbDEMC
hsa-mir-542	unconfirmed	hsa-mir-216a	dbDEMC
hsa-mir-370	HMDD; dbDEMC	hsa-mir-433	dbDEMC
hsa-mir-212	HMDD; dbDEMC	hsa-mir-134	HMDD; dbDEMC
hsa-mir-30e	HMDD	hsa-mir-376a	HMDD; dbDEMC

(WKNKN) to solve this problem, thus enhancing the prediction performance of our model. Third, as a common model of recommendation systems, NMF also plays a crucial role in bioinformatics. However, standard NMF did not achieve satisfactory performance. Therefore, we added the Tikhonov (L_2), graph Laplacian regularization terms and the $L_{2,1}$-norm into the standard NMF, which makes this model more reliable and robust. Finally, the AUCs of $GRL_{2,1}$-NMF are higher than those of some excellent models.

Note that DNSGRMF [53], which also predicts miRNA-disease connections, is a graph regularized method similar to $GRL_{2,1}$-NMF. Both methods decompose the original matrix Y into two matrices W and H, and then we can acquire a recovery matrix $Y^* = W * H$. It is worth noting that $GRL_{2,1}$-NMF is based on non-negative factorization, while DNSGRMF is based on graph regularized matrix factorization. DNSGRMF has no constraints, while $GRL_{2,1}$-NMF has two constraints of $W \geq 0$ and $H \geq 0$.

Nevertheless, our model still has room for improvement. First, miRNA information and disease information did not integrate perfectly, and we will improve this in future studies. Second, there may be more appropriate regularization terms that can improve the performance for miRNA-disease association prediction.

Conclusions

It is meaningful and significant to predict disease-related miRNAs in studying the intrinsic aetiological factors of human diseases. A new model named $GRL_{2,1}$-NMF was developed in this work for potential miRNA-disease association prediction. First, we integrated experimentally validated connections between miRNAs and disease as well as miRNA functional similarities along with two kinds of disease semantic similarities, and then we calculated the GIP kernel similarities of microRNAs and diseases. Moreover, we used WKNKN to convert the value of matrix Y into a decimal between 0 and 1 and decrease the sparsity of matrix Y. Furthermore, the Tikhonov

(L_2), graph Laplacian regularization terms and the $L_{2,1}$-norm were added into the traditional NMF model for predicting miRNA-disease connections. In addition, the Tikhonov regularization was utilized to penalize the non-smoothness of W and H, and the graph Laplacian regularization was primarily intended to guarantee local-based representation by leveraging the geometric structure of the data. The $L_{2,1}$-norm was added to increase the disease matrix sparsity and eliminate unattached disease pairs.

Our method performs well in global LOOCV, 5-CV and case studies in heterogeneous omics data. The experimental results indicate that $GRL_{2,1}$-NMF can effectively and powerfully infer disease-related miRNAs, even if there are no known miRNA-disease associations. However, this method still has limitations that need further research. First, our similarity measurement for $GRL_{2,1}$-NMF might not be perfect, and other miRNA information still needs to be taken into account. Moreover, there is still room for improvement in the predictive performance of our method.

Methods

Human miRNA-disease associations

We collected information on all experimentally validated human miRNA-disease associations stored in the HMDD v2.0 database [10]. An adjacency matrix $Y \in R^{n \times m}$ was established to represent the manually verified human miRNA-disease associations, and the rows and columns of matrix Y represent miRNA m_i interactions and diseases d_j interactions, respectively. Therefore, in this study, the number of rows and columns in Y was 495 and 383, respectively. If a miRNA m_i has a known connection with a disease d_j, $Y_{ij} = 1$, else $Y_{ij} = 0$.

MiRNA functional similarity

There is a hypothesis that if two miRNAs are similar functionally, they are more likely to have connections with diseases that have high similarity, and vice versa [61, 62]. Wang et al. [63] shared their investigation results, and researchers can download miRNA functional similarity information at http://www.cuilab.cn/files/images/cuilab/misim.zip. Here, we established a matrix S^m that was denoted as the microRNA functional similarities. The item $S^m(m_i, m_j)$ denotes the functional similarities among microRNAs m_i and m_j.

Disease semantic similarity method 1

In this study, we take full advantage of the hierarchical directed acyclic graphs (DAGs) for disease similarity measurement based on the strategy of Wang et al. [63], and the disease DAG could be downloaded from the Medical Subject Headings (MeSH) database. $DAG_d = (d, T_d, E_d)$ denotes the hierarchical DAG of disease d, where T_d denotes the disease collection, and E_d denotes links set in the DAG. According to the DAGs, the semantic values of disease D can be computed as Eq. (4).

$$DV1(D) = \sum_{d \in T(D)} D1_D(d) \tag{4}$$

where $D1_D(d)$ denotes the semantic contributions of disease d' to disease d, and Δ denotes the semantic contribution factor ($\Delta = 0.5$) [63].

$$\begin{cases} D1_D(d) = 1 \ if \ d = D \\ D1_D(d) = \max\left\{\Delta * D1_D\left(d'\right) | d' \in child \ of \ d\right\} if \ d \neq D \end{cases} \tag{5}$$

Therefore, two diseases would likely have greater similarities if they share a larger part of their DAGs, and we can calculate semantic similarities between disease d_i and d_j as follows:

$$S_1^d(d_i, d_j) = \frac{\sum_{t \in T(d_i) \cap T(d_j)} \left(D1_{d_i}(t) + D1_{d_j}(t)\right)}{DV1(d_i) + DV1(d_j)} \tag{6}$$

Disease semantic similarity method 2

In the strategy for calculating disease semantic similarities above, diseases that shared one layer of DAG_d shared a common contribution value. However, if some diseases merely exist in fewer DAGs, then these diseases are called more specific diseases and should have a higher semantic contribution to disease d. In view of the algorithm presented by [19, 45], we can calculate the semantic contributions of disease d to disease D and the semantic values of disease D as Eq. (7) and Eq. (8), respectively.

$$D2_D(d) = -\log\left(\frac{the \ number \ of \ DAGs \ including \ d}{the \ number \ of \ diseases}\right) \tag{7}$$

$$DV2(D) = \sum_{d \in T(D)} D2_D(d) \tag{8}$$

where d denotes any investigated disease. Finally, we could calculate the semantic similarities of diseases d_i and d_j as Eq. (9).

$$S_2^d(d_i, d_j) = \frac{\sum_{t \in T(d_i) \cap T(d_j)} \left(D2_{d_i}(t) + D2_{d_j}(t)\right)}{DV2(d_i) + DV2(d_j)} \tag{9}$$

where the numerator of Equation (9) represents the common ancestor nodes of diseases d_i and d_j, and the denominator denotes the entire ancestor nodes of diseases d_i and d_j.

Gaussian interaction profile kernel similarity for diseases and MiRNAs

If two diseases are similar, they are likely to have associations with microRNAs that are functionally approximate, and vice versa [61–64]. Gaussian interaction profile (GIP) kernel similarities have been adopted to quantify disease similarities and miRNA similarities [60, 65, 66]. We also calculated GIP kernel similarities for diseases and miRNAs in this work. First, based on whether disease $d_i(m_j)$ has a known connection with each miRNA (disease) of the adjacency matrix Y, the interaction profiles $IP(d_i)$ and $IP(m_j)$ were constructed for disease d_i and miRNA m_j, respectively. Then, the GIP kernel similarity between a disease pair and a miRNA pair is computed as Equation (10) and Equation (11), respectively.

$$GD(d_i, d_j) = \exp\left(-\beta_d \left\| IP(d_i) - IP(d_j) \right\|^2\right) \tag{10}$$

$$GM(m_i, m_j) = \exp\left(-\beta_m \left\| IP(m_i) - IP(m_j) \right\|^2\right) \tag{11}$$

Here, the kernel bandwidths β_m and β_d are described as Equation (12) and Equation (13), respectively, where $\beta_m^{'}$ and $\beta_m^{'}$ are both the original bandwidths.

$$\beta_m = \beta_m^{'} / \frac{1}{nm} \sum_{i=1}^{nm} \left\| IP(m_i) \right\|^2 \tag{12}$$

$$\beta_d = \beta_d^{'} / \frac{1}{nd} \sum_{i=1}^{nd} \left\| IP(d_i) \right\|^2 \tag{13}$$

In summary, the matrix GD and GM denote the GIP kernel similarity for diseases and miRNAs, respectively.

Integrated similarity for diseases and MiRNAs

According to the various similarity measurement methods mentioned above, we combined the GIP kernel similarities with two disease semantic similarities as well as the miRNA functional similarities to obtain integrated disease similarities and integrated miRNA similarities, respectively. The weight setting problem of the above similarities is described in detail in the Results section, and we chose the following measurement strategy according to the experimental results. Specifically, if two miRNAs m_i and m_j had functional similarities, then the final similarity was the functional similarity. If two miRNAs m_i and m_j did not have functional similarities, then the final similarity was the GIP kernel similarity. Hence, the miRNA similarities score matrix SM between miRNA m_i and miRNA m_j is established as follows. Similarly, the disease similarity matrix SD is computed as follows:

$$SM(m_i, m_j) = \begin{cases} S^m(m_i, m_j) & m_i \text{ and } m_j \text{ have functional similarity} \\ GM(m_i, m_j) & \text{otherwise} \end{cases} \tag{14}$$

$$SD(d_i, d_j) = \begin{cases} \dfrac{S_1^d(d_i, d_j) + S_2^d(d_i, d_j)}{2} & d_i \text{ and } d_j \text{ have semantic similarity} \\ GD(d_i, d_j) & \text{otherwise} \end{cases} \tag{15}$$

Weighted K nearest known Neighbours (WKNKN) for MiRNAs and diseases

Let $M = \{m_1, m_2, \ldots, m_n\}$ and $D = \{d_1, d_2, \ldots, d_m\}$ represent the collection of n microRNAs and m diseases, respectively. We described the quantity of the investigated miRNAs and diseases as n and m, respectively, and then established an association matrix $Y \in R^{n \times m}$ to denote the known human microRNA-disease connections according to the HMDD v2.0 [10] database. If a miRNA m_i had been manually validated to be related to a disease d_j, then Y_{ij} is equal to 1; otherwise, it is equal to 0. $Y(m_i) = \{Y_{i1}, Y_{i2}, \ldots, Y_{im}\}$, namely, the ith row vector of matrix Y, represents the interaction profile for miRNA m_i. Similarly, $Y(d_j) = \{Y_{1j}, Y_{2j}, \ldots, Y_{nj}\}$, the jth column vector of matrix Y, represents the interaction profile for disease d_j. In this study, we investigated 495(n) miRNAs and 383(m) diseases, yet the adjacency matrix $Y \in R^{n \times m}$ has merely 5430 known entries; thus, Y is a sparse matrix. Here, we performed a pre-processing procedure named weighted K nearest known neighbours (WKNKN) [50] for miRNAs and diseases without any known associations to resolve the abovementioned sparse problem and thus improve the prediction accuracy. After executing WKNKN, the entry Y_{ij} was replaced with a continuous value ranging from 0 to 1, and the specific steps are as follows.

First, we acquired the interaction profile of each miRNA m_q according to the functional similarity between m_q and its K nearest known miRNAs as follows:

$$Y_m(m_q) = \frac{1}{Q_m} \sum_{i=1}^{K} w_i Y(m_i) \tag{16}$$

where m_1 to m_K are the miRNAs sorted in descending order based on their similarities to m_q; w_i is the weight factor, and $w_i = \alpha^{i-1} * S^m(m_i, m_q)$; in other words, the higher the similarity between m_i and m_q is, the higher the weight. $\alpha \in [0, 1]$ is a decay term, and $Q_m = \sum_{1 \le i \le K} S^m(m_i, m_q)$ is the normalization coefficient.

Second, we acquired the interaction profile of each miRNA d_p according to the semantic similarity between d_p and its K nearest known diseases as follows:

$$Y_d(d_p) = \frac{1}{Q_d}\sum_{j=1}^{K} w_j Y(d_j) \tag{17}$$

where d_1 to d_K are the diseases sorted in descending order based on their similarities to d_p; w_j is the weight factor, and $w_j = \alpha^{j-1} * S^d(d_j, d_p)$; in other words, the higher the similarity between d_j and d_p is, the higher weight. $Q_d = \Sigma_{1 \le j \le K} S^d(d_j, d_p)$ is the normalization term.

Finally, we took the average of the above two values instead of $Y_{ij} = 0$, indicating the overall likelihood of the interaction between m_i and d_j. Then, we integrated the above two matrices Y_m and Y_d acquired from different datasets, replaced $Y_{ij} = 0$ with the related likelihood scores, and then updated the original adjacency matrix Y as follows:

$$Y_{md} = a_1 Y_m + a_2 Y_d / \sum a_i (i = 1, 2) \tag{18}$$

$$Y = \max(Y, Y_{md}) \tag{19}$$

where a_i is the weight coefficient and $a_1 = a_2 = 1$.

Standard NMF

In recent years, as one of the common methods of recommendation systems, nonnegative matrix factorization (NMF) has been widely used as an effective prediction algorithm in the field of bioinformatics [67, 68]. Two non-negative matrices W and H, which are optimal approximations to the original matrix Y, can be found by NMF, where W and H satisfy Equation (20).

$$\mathrm{Y} \approx \mathrm{W}H^T, \mathrm{s.t.}\, W \ge 0, H \ge 0 \tag{20}$$

In this work, matrix $Y \in R^{n \times m}$ was used to represent the known miRNA-disease associations, and NMF can decompose this matrix into two matrices, namely, $W \in R^{n \times k}$ and $H \in R^{m \times k}$. Here, we express the question of the miRNA-disease association identification problem as the objective function (Equation (21)).

$$min_{W,H} \left\| Y - WH^T \right\|_F^2 \mathrm{s.t.} \mathrm{W} \ge 0, \mathrm{H} \ge 0 \tag{21}$$

where $\|\cdot\|_F$ represents the Frobenius norm of a matrix. Equation (21) can be optimized by taking advantage of the iterative update algorithm presented by [69].

However, standard NMF does not ensure the sparsity of decomposition; therefore, local-based representations are not always generated [70, 71]. Some researchers have developed sparse constraints on NMF [46–48].

$GRL_{2,1}$-NMF

Here, a new nonnegative matrix factorization method was presented to identify underlying miRNA-disease connections. The flow chart of $GRL_{2,1}$-NMF is shown in Fig. 3. We incorporated Tikhonov (L_2), graph Laplacian regularization terms and the $L_{2,1}$-norm into the traditional NMF model for predicting miRNA-disease connections. The Tikhonov regularization is utilized to penalize the non-smoothness of W and H [48, 54, 55], and the graph Laplacian regularization is primarily intended to ensure local-based representation by

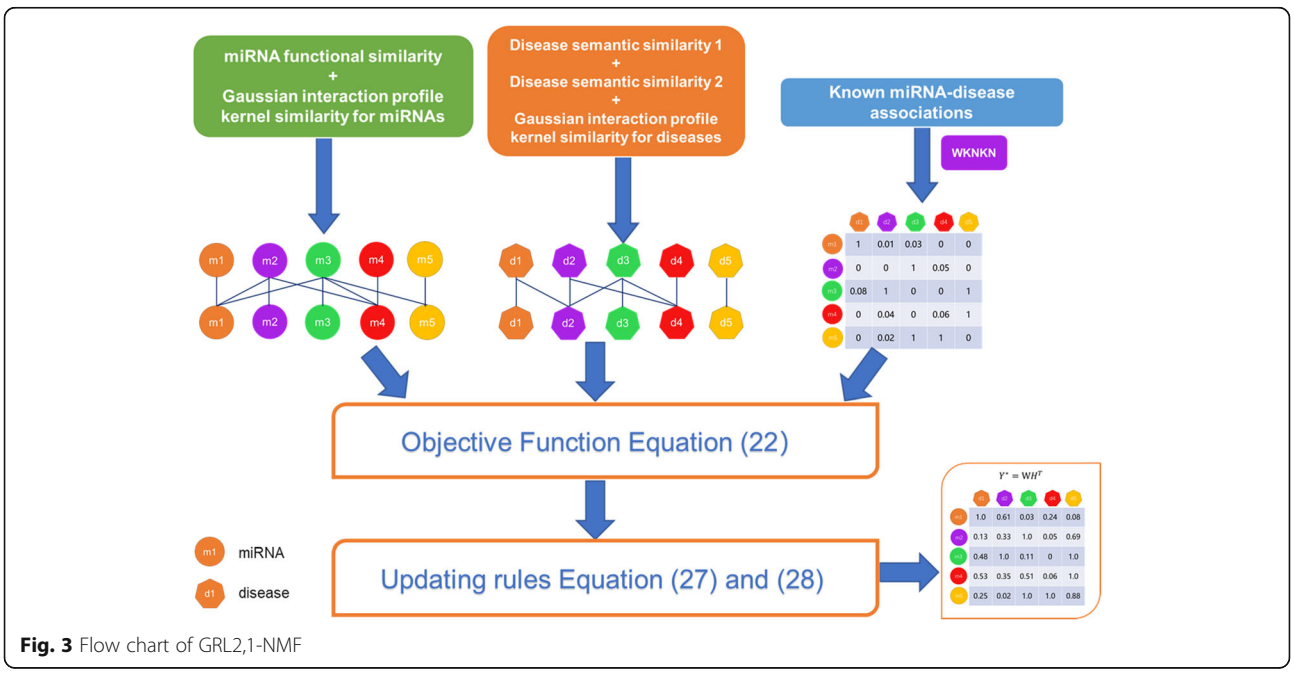

Fig. 3 Flow chart of GRL2,1-NMF

leveraging the geometric structure of the data [56]. The $L_{2,\ 1}$-norm was added to increase the disease matrix sparsity and eliminate unattached disease pairs [30, 52, 53]. The optimization problem of $GRL_{2,\ 1}$-NMF can be formularized as follows:

$$\begin{aligned} min_{W,H} \left\| Y - WH^T \right\|_F^2 + \lambda_l \left(\|W\|_F^2 + \|H\|_F^2 \right) + \lambda_l \|H\|_{2,1} \\ + \lambda_m Tr\left(W^T L_m W \right) + \lambda_d Tr\left(H^T L_D H \right) \\ s.t. W \geq 0, H \geq 0 \end{aligned} \tag{22}$$

where $\|\cdot\|_F$ represents the Frobenius norm of a matrix; $\|\cdot\|_{2,\ 1}$ represents the $L_{2,\ 1}$-norm; Tr(·) denotes the trace of a matrix; and λ_l, λ_m and λ_d are regularization coefficients. Let S^m and S^d be miRNA and disease similarity networks; and let D_m and D_d be the diagonal matrices whose elements are row element or column element sums of S^m and S^d respectively. We define $L_m = D_m - S^m$ and $L_d = D_d - S^d$ as the graph Laplacian matrices for S^m and S^d [72], respectively; the first item denotes the similar matrix of the model for the purpose of searching for the matrices W and H. The next term is the Tikhonov regularization. The third item introduces the $L_{2,\ 1}$-norm into matrix H. The last two items refer to the graph regularization of microRNAs and diseases.

Optimization

Considering the two nonnegative constraints of the objective function, namely, $W \geq 0$ and $H \geq 0$, we utilized Lagrange multipliers to address the optimization problem in Equation (22). First, the Lagrange function L_f is as follows:

$$\begin{aligned} L_f = Tr\left(YY^T \right) - 2Tr\left(YHW^T \right) + Tr\left(WH^T HW^T \right) \\ + \lambda_l Tr\left(WW^T \right) + \lambda_l Tr\left(HH^T \right) + \lambda_l \|H\|_{2,1} \\ + \lambda_m Tr\left(W^T L_m W \right) + \lambda_d Tr\left(H^T L_d D \right) H \\ + Tr\left(\varnothing W^T \right) + Tr\left(\phi H^T \right) \end{aligned} \tag{23}$$

The partial derivatives of the above functions L_f for W and H are:

$$\begin{aligned} \frac{\partial L_f}{\partial W} = -2YH + 2WH^T H + 2\lambda_l W + 2\lambda_m L_m W \\ + \varnothing \end{aligned} \tag{24}$$

$$\begin{aligned} \frac{\partial L_f}{\partial H} = -2Y^T W + 2HW^T W + 2\lambda_l H + 2\lambda_l AH \\ + 2\lambda_d L_d \mathrm{H} + \phi \end{aligned} \tag{25}$$

where A is a diagonal matrix, and the formula is as follows:

$$[A]_{i,j} = 1 \Big/ \|H^s\|_2 = 1 \Big/ \sqrt{\sum_{j=1}^{m} |H_{s,j}|^2} \tag{26}$$

Therefore, we obtained the updating rules expressed as Equations (27) and (28):

$$w_{ik} \leftarrow w_{ik} \frac{(YH + \lambda_m S^m W)_{ik}}{\left(WH^T H + \lambda_l W + \lambda_m D_m W \right)_{ik}} \tag{27}$$

$$h_{ik} \leftarrow h_{ik} \frac{\left(Y^T W + \lambda_d S^d H \right)_{ik}}{\left(HW^T W + \lambda_l H + \lambda_l AH + \lambda_d D_d H \right)_{ik}} \tag{28}$$

According to Equation (27) and Equation (28), the nonnegative matrices W and H are updated until convergence. Eventually, we obtained a matrix of $Y^* = WH^T$, which is based on interactions among microRNAs and disease. We ranked predicted disease-connected miRNAs according to the elements in matrix Y^*. In theory, the higher-ranking miRNAs in each column of Y^* tend to be connected with the matching disease.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12859-020-3409-x.

Additional file 1. disease semantic similarity1.txt. This is disease semantic similarity 1, which integrated 383 disease semantic similarities.

Additional file 2. disease semantic similarity2.txt. This is disease semantic similarity 2, which integrated 383 disease semantic similarities.

Additional file 3. miRNA functional similarity.txt. This is the miRNA functional similarity, which integrated 495 miRNA functional semantic similarities.

Additional file 4. knownassociation.txt. This is a known miRNA-disease association matrix that was downloaded from HMDD v2.0. It includes 5430 known miRNA-disease associations between 495 miRNAs and 383 diseases.

Additional file 5. diseases_list.xlsx. This file lists 383 disease names.

Additional file 6. miRNAs_list.xlsx. This file lists 495 miRNA names.

Abbreviations

5-CV: Five-fold cross validation; AUC: The area under the ROC curve; DAG: Directed acyclic graph; dbDEMC: Database of differentially expressed miRNAs in human cancers; FPR: False positive rate; GIP: GAUSSIAN interaction profiles; HMDD: Human microRNA disease database; LOOCV: Leave-one-out cross validation; miRNA: MicroRNA; NMF: Nonnegative matrix factorization; ROC: Receiver operating characteristic; TPR: True positive rate; WKNKN: Weighted K nearest known neighbours

Acknowledgements

Not applicable.

Authors' contributions

ZG and YTW collected the data. ZG, YTW, QWW, JCN and CHZ conceived and designed the experiments. ZG implemented the experiments. ZG and CHZ analysed the results. ZG and CHZ wrote the paper. All authors read and approved the final manuscript.

Funding

This work was supported by the National Natural Science Foundation of China (Nos. U19A2064, 61873001, 61872220, 61672037, 61861146002 and

61732012). The funding bodies did not play any role in the design of the study or collection, analysis and interpretation of data or in writing the manuscript.

Availability of data and materials
The dataset(s) supporting the conclusions of this article is (are) included within the article (and its additional files).

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 28 August 2019 Accepted: 11 February 2020
Published online: 18 February 2020

References

1. Victor A. microRNAs: tiny regulators with great potential. Cell. 2001;107:823–6.
2. Ambros V. The functions of animal microRNAs. Nature. 2004;431:350–5.
3. Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell. 2004;116:281–97.
4. Lee RC, Feinbaum RL, Ambros V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell. 1993;75:843–54.
5. Jopling CL, Yi M, Lancaster AM, Lemon SM, Sarnow P. Modulation of hepatitis C virus RNA abundance by a liver-specific MicroRNA. Science. 2005;309:1577–81.
6. Kozomara A, Griffiths-Jones S. miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res. 2011;39:D152–7.
7. Bartel DP. MicroRNAs. Target recognition and regulatory functions. Cell. 2009;136:215–33.
8. Harfe BD. MicroRNAs in vertebrate development. Curr Opin Genet Dev. 2005;15:410–5.
9. Chou CH, Chang NW, Shrestha S, Hsu SD, Lin YL, et al. miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. Nucleic Acids Res. 2016;44:D239–47.
10. Li Y, Qiu C, Tu J, Geng B, Yang J, Jiang T, Cui Q. HMDD v2.0: a database for experimentally supported human microRNA and disease associations. Nucleic Acids Res. 2014;42:D1070–4.
11. Meola N, Gennarino VA, Banfi S. microRNAs and genetic diseases. Pathogenetics. 2009;2:7.
12. Calin GA, Croce CM. MicroRNA signatures in human cancers. Nat Rev Cancer. 2006;6:857–66.
13. Luo J, Xiao Q. A novel approach for predicting microRNA-disease associations by unbalanced bi-random walk on heterogeneous network. J Biomed Inform. 2017;66:194–203.
14. Zeng X, Zhang X, Zou Q. Integrative approaches for predicting microRNA function and prioritizing disease-related microRNA using biological interaction networks. Brief Bioinform. 2016;17:193–203.
15. Chen X, Xie D, Zhao Q, You ZH. MicroRNAs and complex diseases: from experimental results to computational models. Brief Bioinform. 2019;20:515–39.
16. Jiang Q, Hao Y, Wang G. Prioritization of disease microRNAs through a human phenome-microRNAome network. BMC Syst Biol. 2010;4:S2.
17. Li X, Wang Q, Zheng Y, Lv S, Ning S, Sun J, Huang T, Zheng Q, Ren H, Xu J, et al. Prioritizing human cancer microRNAs based on genes' functional consistency between microRNA and cancer. Nucleic Acids Res. 2011;39:e153.
18. Xu C, Ping Y, Li X, Zhao H, Wang L, Fan H, Xiao Y, Li X. Prioritizing candidate disease miRNAs by integrating phenotype associations of multiple diseases with matched miRNA and mRNA expression profiles. Mol BioSyst. 2014;10:2800–9.
19. Xuan P, Han K, Guo M, Guo Y, Li J, Ding J, Liu Y, Dai Q, Li J, Teng Z, et al. Prediction of microRNAs associated with human diseases based on weighted k most similar neighbors. PLoS One. 2013;8:e70204.
20. Mork S, Pletscher-Frankild S, Palleja Caro A, Gorodkin J, Jensen LJ. Protein-driven inference of miRNA-disease associations. Bioinformatics. 2014;30:392–7.
21. Chen X, Yan CC, Zhang X, You ZH, Deng L, Liu Y, Zhang Y, Dai Q. WBSMDA: within and between score for MiRNA-disease association prediction. Sci Rep. 2016;6:21106.
22. Chen X, Yan CC, Zhang X, You ZH, Huang YA, GY Y. HGIMDA: heterogeneous graph inference for miRNA-disease association prediction. Oncotarget. 2016;7(40):65257–69.
23. Yu H, Chen X, Lu L. Large-scale prediction of microRNA-disease associations by combinatorial prioritization algorithm. Sci Rep. 2017;7:43792.
24. Chen X, Wu QF, Yan GY. RKNNMDA: ranking-based KNN for MiRNA-disease association prediction. RNA Biol. 2017;14:952–62.
25. Chen X, Cheng JY, Yin J. Predicting microRNA-disease associations using bipartite local models and hubness-aware regression. RNA Biol. 2018;15:1192–205.
26. You Z-H, Huang Z-A, Zhu Z, Yan G-Y, Li Z-W, Wen Z, Chen X. PBMDA: a novel and effective path-based computational model for miRNA-disease association prediction. PLoS Comput Biol. 2017;13:e1005455.
27. Chen X, Liu MX, Yan GY. RWRMDA: predicting novel human microRNA-disease associations. Mol BioSyst. 2012;8:2792–8.
28. Shi H, Xu J, Zhang G. Walking the interactome to identify human miRNA-disease associations through the functional link between miRNA targets and disease genes. BMC Syst Biol. 2013;7:101.
29. Xuan P, Han K, Guo Y, Li J, Li X, Zhong Y, Zhang Z, Ding J. Prediction of potential disease-associated microRNAs based on random walk. Bioinformatics. 2015;31:1805–15.
30. Liu Y, Zeng X, He Z, Zou Q. Inferring microRNA-disease associations by random walk on a heterogeneous network with multiple data sources. IEEE/ACM Trans Comput Biol Bioinform. 2017;14:905–15.
31. Wong L, You ZH, Ming Z, Li J, Chen X, Huang YA. Detection of interactions between proteins through rotation Forest and local phase quantization descriptors. Int J Mol Sci. 2016;17:21.
32. Xu J, Li CX, Lv JY, Li YS, Xiao Y, Shao TT, Huo X, Li X, Zou Y, Han QL, et al. Prioritizing candidate disease miRNAs by topological features in the miRNA target-dysregulated network: case study of prostate cancer. Mol Cancer Ther. 2011;10:1857–66.
33. Chen X, Yan GY. Semi-supervised learning for potential human microRNA-disease associations inference. Sci Rep. 2014;4:5501.
34. Li JQ, Rong ZH, Chen X. MCMDA: matrix completion for MiRNA-disease association prediction. Oncotarget. 2017;8:21187–99.
35. Chen X, Wang CC, Yin J, You ZH. Novel human miRNA-disease association inference based on random Forest. Mol Ther Nucleic Acids. 2018;13:568–79.
36. Chen X, Zhang DH, You ZH. A heterogeneous label propagation approach to explore the potential associations between miRNA and disease. J Transl Med. 2018;16:348.
37. Chen X, Yan CC, Zhang X, Li Z, Deng L, Zhang Y, Dai Q. RBMMMDA: predicting multiple types of disease-microRNA associations. Sci Rep. 2015;5:13877.
38. Chen X, Gong Y, Zhang DH, You ZH, Li ZW. DRMDA: deep representations-based miRNA-disease association prediction. J Cell Mol Med. 2018;22:472–85.
39. Chen X, Huang L, Xie D, Zhao Q. EGBMMDA: extreme gradient boosting machine for MiRNA-disease association prediction. Cell Death Dis. 2018;9:3.
40. Chen X, Zhu C-C, Yin J. Ensemble of decision tree reveals potential miRNA-disease associations. PLOS Comput Biol. 2019;15:e1007209.
41. Chen X, Huang L. LRSSLMDA: Laplacian regularized sparse subspace learning for MiRNA-disease association prediction. PLoS Comput Biol. 2017;13:e1005912.
42. Zhao Y, Chen X, Yin J. Adaptive boosting-based computational model for predicting potential miRNA-disease associations. Bioinformatics. 2019;35:4730–8.
43. Zou Q, Li J, Hong Q, Lin Z, Wu Y, Shi H, Ju Y. Prediction of MicroRNA-disease associations based on social network analysis methods. Biomed Res Int. 2015;2015:810514.
44. Chen X, Zhou Z, Zhao Y. ELLPMDA: ensemble learning and link prediction for miRNA-disease association prediction. RNA Biol. 2018;15:807–18.
45. Chen X, Xie D, Wang L, Zhao Q, You ZH, Liu H. BNPMDA: bipartite network projection for MiRNA-disease association prediction. Bioinformatics. 2018;34:3178–86.

46. Zhao Y, Chen X, Yin J. A novel computational method for the identification of potential miRNA-disease association based on symmetric non-negative matrix factorization and Kronecker regularized Least Square. Front Genet. 2018;9:324.
47. Zhong Y, Xuan P, Wang X, Zhang T, Li J. A non-negative matrix factorization based method for predicting disease-associated miRNAs in miRNA-disease bilayer network. Bioinformatics. 2018;34:267–77.
48. Xiao Q, Luo J, Liang C, Cai J, Ding P. A graph regularized non-negative matrix factorization method for identifying microRNA-disease associations. Bioinformatics. 2018;34:239–48.
49. Chen X, Yin J, Qu J, Huang L. MDHGI: matrix decomposition and heterogeneous graph inference for miRNA-disease association prediction. PLoS Comput Biol. 2018;14:e1006418.
50. Ezzat A, Zhao PL. Drug-target interaction prediction with graph regularized matrix factorization. IEEE/ACM Transactions On Computational Biology And Bioinformatics. 2017;14:646–56.
51. Liu JX, Wang D, Gao YL, Zheng CH, Shang JL, Liu F, Xu Y. A joint-L2,1-norm-constraint-based semi-supervised feature extraction for RNA-Seq data analysis. Neurocomputing. 2017;228:263–9.
52. Cui Z, Gao YL, Liu JX, Wang J, Shang J, Dai LY. The computational prediction of drug-disease interactions using the dual-network L2,1-CMF method. BMC Bioinformatics. 2019;20:5.
53. Gao MM, Cui Z, Gao YL, Liu JX, Zheng CH. Dual-network sparse graph regularized matrix factorization for predicting miRNA-disease associations. Mol Omics. 2019;15:130–7.
54. Pauca VP, Shahnaz F, Berry MW. Text mining using non-negative matrix factorization. In: Proceedings of the 2004 SIAM International Conference on Data Mining Society for Industrial and Applied Mathematics; 2004. p. 452–6.
55. Guan N, Tao D, Luo Z, Yuan B. Manifold regularized discriminative nonnegative matrix factorization with fast gradient descent. IEEE Trans Image Process. 2011;20:2030–48.
56. Cai D, He X, Han J. Graph regularized non-negative matrix factorization for data representation. IEEE Trans Pattern Anal Mach Intell. 2011;33:1548–60.
57. Yang Z, Wu L, Wang A, Tang W, Zhao Y, Zhao H, AEJNar T. dbDEMC 2.0: updated database of differentially expressed miRNAs in human cancers. Nucleic Acids Res. 2016;45:D812–8.
58. Jiang Y, Liu B, Yu L, Yan C, Bian H. Predict MiRNA-disease association with collaborative filtering. Neuroinformatics. 2018;16:363–72.
59. Shao B, Liu B, Yan C. SACMDA: MiRNA-disease association prediction with short acyclic connections in heterogeneous graph. Neuroinformatics. 2018;16:373–82.
60. Chen X, Wang L, Qu J, Guan NN, Li JQ. Predicting miRNA-disease association based on inductive matrix completion. Bioinformatics. 2018;34:4256–65.
61. Goh KI, Cusick ME, Valle D. The human disease network. Proc Natl Acad Sci. 2007;104:8685–90.
62. Lu M, Zhang Q, Deng M. An analysis of human microRNA and disease associations. PLoS One. 2008;3:e3420.
63. Wang D, Wang JY, Lu M. Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. Bioinformatics. 2010;26:1644–50.
64. Bandyopadhyay S, Mitra R, Maulik U. Development of the human cancer microRNA network. Silence. 2010;1:6.
65. Chen X, Huang YA, You ZH, Yan GY, Wang XS. A novel approach based on KATZ measure to predict associations of human microbiota with non-infectious diseases. Bioinformatics. 2018;34:1440.
66. Chen X, Yan CC, Zhang X, Zhang X, Dai F, Yin J, Zhang Y. Drug-target interaction prediction: databases, web servers and computational models. Brief Bioinform. 2016;17:696–712.
67. Zheng CH, Huang DS, Zhang L, Kong XZ. Tumor clustering using nonnegative matrix factorization with gene selection. IEEE Trans Inf Technol Biomed. 2009;13:599–607.
68. Huang DS, Zheng CH. Independent component analysis-based penalized discriminant method for tumor classification using gene expression data. Bioinformatics. 2006;22:1855–62.
69. Lee DD, Seung HS. Learning the parts of objects by nonnegtive matrix factorization. Nature. 1999;401:788–91.
70. Li X, Cui G, Dong Y. Graph regularized non-negative low-rank matrix factorization for image clustering. IEEE Trans Cybern. 2017;47:3840–53.
71. Wang JY, Almasri I, Gao X. Adaptive Graph Regularized Nonnegative Matrix Factorization via Feature Selection. In: Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012) IEEE; 2012. p. 963–6.
72. Liu X, Zhai D, Zhao D, Zhai G, Gao W. Progressive image denoising through hybrid graph Laplacian regularization: a unified framework. IEEE Trans Image Process. 2014;23:1491–503.

Publisher's Note