Journal of Biomedical Informatics 102 (2020) 103358

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

IMIPMF: Inferring miRNA-disease interactions using probabilistic matrix factorization

Jihwan Ha[a], Chihyun Park[a], Chanyoung Park[b], Sanghyun Park[a,*]

[a] Department of Computer Science, Yonsei University, 134 Sinchon-dong, Seodaemun-gu, Seoul, South Korea
[b] Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, USA

ARTICLE INFO

Keywords:
miRNA
Disease
miRNA–disease association
Probabilistic matrix factorization

ABSTRACT

Recently, increasing evidence have reported that microRNAs (miRNAs) play key roles in a variety of biological processes. Therefore, the identification of novel miRNA–disease associations can shed new light on disease etiology and pathogenesis. Till now, various computational methods have been proposed to predict potential miRNA–disease associations by reducing the experimental costs and time consumption. However, most existing methods are highly dependent on known miRNA–disease associations. Therefore, the prediction of new miRNAs (i.e., miRNAs without known associated diseases) and new diseases (i.e., diseases without known associated miRNAs) has become challenging. In this paper, we present IMIPMF, a novel method for predicting miRNA–disease associations using probabilistic matrix factorization (PMF), which is a machine learning technique that is widely used in recommender systems. Predicting the rating scores that a user may assign to each item in a recommender system is analogous to predicting miRNA–disease associations. By applying PMF, our model not only identifies novel miRNA-disease associations, but also overcomes the common problem of incompatibility with miRNAs without any known associated disease, which was a limitation of most previous computational methods. We demonstrated that our proposed model achieved a high performance with a reliable AUC value of 0.891 by performing 5-fold cross-validation. Overall, IMIPMF is a high-performance machine-learning-based model for predicting miRNA–disease associations, although it only considers known miRNA–disease associations and miRNA expression data.

1. Introduction

MicroRNAs (miRNAs) are endogenous, single-stranded, non-coding RNAs that suppress the expression of target messenger RNAs (mRNAs) at the post-transcriptional level by binding to 3′ untranslated regions (UTRs) [1,2]. However, various studies have demonstrated that miRNAs also function as positive regulators at the post-transcriptional level, directly inducing the pathogenesis of disease mechanisms. Various studies have proved that miRNAs play significant roles in multiple biological processes, such as aging [3,4], apoptosis [5], cell development [6], differentiation [7], metabolism [3], proliferation [8], transduction [9], and viral infection [10]. In the light of this issue, miRNAs have received considerable attention because they have significant impact on disease emergence. Because miRNAs have been found to be related to multiple diseases either directly or indirectly through various biological experiments, it is necessary to trace the relationship between miRNAs and diseases. Various studies have demonstrated the critical role of miRNAs in disease incidence. For example, miR-101, by targeting Stathmin1, was found to be a significant causal factor in breast cancer [11]. Furthermore, miR-143 and miR-145 have been found to be involved in down-regulating colorectal tumors and breast carcinomas [12]. Wang et al. also determined that miR-185, by targeting Vegfa, is one of the main factors involved in breast cancer [13]. Further, miR-355 and miR-31 play significant roles in inhibiting breast cancer [14–16].

Because miRNAs are known to be a significant disease-related factor in various biological processes, identifying the interactions between miRNAs and diseases has become a crucial problem. Several computational methods have been proposed to infer miRNA–disease associations under the assumption that functionally similar miRNAs are inclined to have associations with phenotypically similar diseases [17–19]. Computational methods have also been applied to reduce the experimental cost and time consumption to decrease the overall biological experimental workload. Most conventional methods used for

Abbreviations: miRNA, microRNA; PMF, probabilistic matrix factorization; EF, environmental factor

* Corresponding author.
E-mail addresses: jihwanha@yonsei.ac.kr (J. Ha), chihyun.park@yonsei.ac.kr (C. Park), pcy1302@illinois.edu (C. Park), sanghyun@yonsei.ac.kr (S. Park).

https://doi.org/10.1016/j.jbi.2019.103358
Received 14 June 2019; Received in revised form 11 November 2019; Accepted 12 December 2019
Available online 16 December 2019

predicting miRNA–disease associations rely on a similarity-based approach. Jiang et al. developed a hypergeometric-distribution-based approach to infer miRNA–disease associations using three heterogeneous similarity networks: a human phenome-miRNAome network, disease similarity network, and miRNA functional similarity network. [20]. The main drawback of this method is that it is highly dependent on local neighborhood information, which leads to high false positive and negative rates. This leaves significant room for improvement by utilizing global networks. Chen et al. developed the "random walk with restart for miRNA–disease association" (RWRMDA) algorithm to infer disease-related miRNAs by implementing a random walk algorithm on a miRNA functionality network [21]. Although RWRMDA has shown good predictive accuracy, this method is not suitable for the query diseases without any known related miRNAs. Mørk et al. [22] developed a miRNA-protein-disease (miRPD) approach that utilizes linkages among miRNAs, proteins, and diseases to predict unknown miRNA–-disease associations. In miRPD, they used proteins as the mediators between miRNAs and diseases by integrating the miRNA–protein and protein–disease associations described in the literature. That is to say, the miRNA–disease associations with more shared common proteins have a larger possibility of being involved in disease incidence. Xuan et al. presented an HDMP model that infers miRNA–disease associations by considering the weighted k most similar neighbors of each node [23]. However, owing to the strong dependency on the neighbors of the miRNAs, the prediction of disease without known related miRNAs remains limited. Chen et al. further presented a miRNA–disease prediction model named "heterogeneous graph inference for miRNA–disease association prediction" (HGIMDA) [24]. In HGIMDA, a heterogeneous graph was constructed by combining a human miRNA–disease association network, disease–disease similarity and miRNA–miRNA similarity. They analyzed all three-length paths on the constructed heterogeneous graph and demonstrated an improved performance over those of previous conventional similarity-based models. Chen et al. also proposed a semi-supervised learning-based model known as "regularized least squares for miRNA–disease association" (RLSMDA) [25]. RLSMDA prioritized candidate miRNA–disease associations through a semi-supervised learning framework without negative samples. The critical drawbacks of RLSMDA are the difficulty in selecting the optimal parameters of each classifier and combining classifiers from the different spaces. Chen et al. also proposed another model known as "with and between score for miRNA–disease association prediction" (WBSMDA) [26]. This model applies information related to disease semantic similarity, miRNA functional similarity, and Gaussian interaction kernel similarity to miRNAs and diseases. Li et al. proposed a label propagation-based model with linear neighborhood similarity to identify potential miRNA–disease associations [27]. To summarize, most previous similarity-based approaches are highly dependent on the miRNAs that are already linked with the query disease. That is, these methods cannot be implemented on miRNAs that do not belong to any other nodes in the network.

In the past few years, vast amounts of biological data have been produced rapidly based on the development of high-throughput techniques. These data provide opportunities to discover the underlying roles of miRNAs in various biological activities [28,29] as well as in the detection of the regulatory mechanisms of circular RNA [30], TF-miRNA-gene network based module detection [31], and discovery of interplay with transcription factors [32]. The application of these data with machine-learning-based models successively induced performance enhancements for disease-related miRNA prediction because machine learning approaches are capable of handling complicated biological datasets. Accumulated evidence has shown that environmental factors (EFs) can be essential for regulating miRNAs. Ha et al. developed a similarity-based miRNA network, in which similarity is estimated under the assumption that phenotypically similar miRNAs are inclined to share more common EFs [33]. Environmental factors consist of alcohol, cigarette, and drug use, as well as diet, stress, and exposure to radiation. By implementing a propagation algorithm on the similarity-based miRNA network, Ha et al. extracted candidate miRNA–disease associations. This method can be enhanced further by considering the chemical structures of EFs. Ha et al. also developed miRNA-disease prediction model called PMAMCA based on matrix factorization, which is a well-known machine learning-based model with proven excellent performance in recommender systems. They utilized miRNA expression data for objective function weights to enhance the prediction accuracy. Using only known miRNA–disease interactions and miRNA expression data, their method outperformed previous models from a technical perspective [34]. Chen et al. proposed another computational model known as the "restricted Boltzmann machine for multiple types of miRNA–disease association prediction" (RBMMMDA) [35]. This model is a two-layered graphical model with hidden units. RBMMMDA was used to discover novel disease-related miRNAs and their corresponding association types. Li et al. developed a matrix completion algorithm to infer miRNA–disease association without using negative samples—"matrix completion for miRNA–disease association prediction" (MCMDA) [36]. The primary advantage of MCMDA is that it only requires the miRNA–disease association as input. However, MCMDA cannot be applied to diseases without any known miRNA associations. Furthermore, the identification of optimal parameters for the model remains critical. Chen et al. developed a "hybrid approach for miRNA–disease association prediction" (HAMDA) model for predicting miRNA–disease associations by combining information propagation and the network structure [37]. However, HAMDA extracted miRNAs using only neighbor nodes in the same layer of the miRNAs and diseases rather than fully utilizing the topological characteristics of the subgraphs of heterogeneous networks. Chen et al. also proposed the method of predicting miRNA–disease associations based on inductive matrix complementation with matrix decomposition and heterogeneous graphs (IMCMDA) [38]. This approach discovers not only the known miRNA–disease associations but also the comprehensive similarity for miRNA and disease. They also developed the "matrix decomposition and heterogeneous graph inference" (MDHGI) for miRNA–disease association prediction, which extracts known miRNA–disease associations using the matrix decomposition algorithm [39]. By adopting the matrix decomposition algorithm, they avoided the noises in the original adjacency matrix, which led to performance improvement. They further developed the more comprehensive method of "Laplacian regularized sparse subspace learning for miRNA–disease association prediction" (LRSSLMDA) [40]. In this method, they constructed miRNA/disease statistical and graph theoretic features as the inputs of the model to predict potential disease-related miRNAs. The Laplacian regularization term was used as the cost function. They also proposed the computational model of "bipartite network projection for MDA prediction" (BNPMDA) for identifying novel miRNA–disease associations [41]. This framework built bias ratings based on three networks: the known miRNA–disease association, disease similarity, and miRNA similarity networks. With the application of machine learning, various scientific fields have shown enormous improvement in performance by solving critical research problems. Since then, machine-learning-based prediction models [42–46], including matrix-factorization-based methods [47], are increasingly being proposed to handle problems such as disease miRNA prioritization, drug target prediction, and cancer-type classification. Chen et al. developed a neoteric Bayesian model (KMFMDA) by combining kernel-based nonlinear dimension reduction and matrix factorization for predicting potential miRNA–disease associations [48]. Gao et al. proposed a method of predicting miRNA–disease associations based on "dual network sparse graph regularized matrix factorization" (DNSGRMF) [49]. In this method, the L2,1-norm was used to make up for the sparsity in unknown associations. They presented effective "nearest profile-based collaborative matrix factorization" (NPCMF) to predict novel disease-related miRNAs. The nearest-neighbor information of miRNAs and diseases was used to derive a reliable similarity function for discovering new miRNAs and diseases

[50]. Xiao et al. recently developed a framework known as "graph regularized nonnegative matrix factorization" (GRNMF), which utilizes heterogenous omics data to infer unknown miRNA–disease associations. Their framework exploits weighted gene network and semantic associations between diseases to measure the similarities between miRNAs and diseases [51]. However, most machine-learning methods tend to utilize negative samples or struggle to adjust model parameters. Moreover, some methods only utilize local information instead of global information, which could be an area of further improvement.

In past years, recommender system algorithms have shown promise in various fields to predict users' preferences for specific objects such as e-commerce and movies or music recommendation. Most companies that sell products to users have gained significant profits by adopting recommender systems. There are two main types of recommender systems, namely, memory-based and model-based. Memory-based approaches (collaborative filtering) perform predictions using recommendations from sets of users that are similar to a new user u, who are identified by exploring a user-item matrix. Model-based approaches only store model parameters and do not need to search through a rating matrix. Therefore, model-based approaches have the advantage of fast prediction once the parameters of the model are fine-tuned. Further, matrix factorization is one of the most popularly used approaches in recommender systems. The success of matrix factorization in various domains is based on its solid mathematical foundation [52,53].

In this paper, we propose a novel approach for predicting potential miRNA–disease associations using the machine learning technique—probabilistic matrix factorization (PMF). By considering the fact that miRNAs may be related to multiple diseases, we can enhance the prediction accuracy and resolve the problem of applicability to miRNAs with no previously known disease associations. The proposed IMIPMF method aims to accomplish three main objectives: (i) predicting known miRNA–disease associations, (ii) inferring new miRNA–disease candidates, and (iii) determining the disease phenotype using latent vectors learned by miRNA–disease associations.

The remainder paper is organized as follows. We review the biological mechanisms of miRNAs and previous computational prediction methods in Section 1. In Section 2, we describe the operational principles of the proposed IMIPMF method and explain the fundamental concepts of matrix factorization in recommender systems. We also enumerate the datasets that were used in this study. In Section 3, we present experimental results for various diseases and the thresholds and applications of latent vectors for classifying disease phenotypes. In Section 4, we discuss the proposed approach and present the results in the context of future research.

2. Materials and methods

In this section, we provide a detailed description of the proposed IMIPMF model and enumerate the datasets that were used in this study.

Fig. 1 shows the workflow of IMIPMF. First, for preprocessing, we obtained miRNA–disease associations from the dbDEMC [54], HMDD [55], and miR2Disease [56] databases. To enhance training, we also utilized miRNA expression data from The Cancer Genome Atlas (TCGA) to construct the weight matrix of the model. The miRNA expression value was used only when the corresponding entry of the miRNA–disease matrix is unknown. After formulating a binary association matrix R with miRNA–disease association datasets, we implemented PMF to learn the latent vectors for miRNAs and diseases. The inner product of each miRNA and disease latent vector (i.e., $\widehat{r_{ij}} = \boldsymbol{m}_i^T \boldsymbol{d}_j$) provides a numerical measure for miRNA–disease association. Finally, after learning the latent vectors, we extracted candidate miRNAs based on the assumption that the miRNAs with higher scores have a high probability of being related to a given disease. We prioritized candidate miRNAs based on the scores assigned by IMIPMF.

2.1. Data

2.1.1. Human miRNA–disease association dataset

Multiple databases contain miRNA–disease associations that have been determined through various biological experiments. We obtained human miRNA–disease association datasets using the dbDEMC, HMDD, and miR2Disease databases. dbDEMC v2.0 is an integrated database that provides differentially expressed miRNAs related to human cancers and human miRNA–disease associations. The current version of dbDEMC stores information on 2224 miRNAs and 36 diseases [54]. HMDD v2.0 provides experimentally supported human miRNA–disease associations in the form of 10,369 entries with information on 572 miRNA genes and 378 diseases from 3511 articles [55]. miR2Disease is a manually curated database containing detailed information on miRNA IDs, disease names, miRNA expression patterns, and miRNA–disease associations [56]. We acquired information on 349 miRNAs and 163 diseases from 3273 entries in the miR2Disease database. The operation of merging records from different databases to construct a group with no duplicate entries and unifying the disease name based on MeSH disease terms were performed.

2.1.2. miRNA expression dataset

Because several types of meaningful biological data have been generated with the development of high-throughput techniques, we adopted miRNA expression data for the weight matrix W of the cost function to reduce the effects of low values in the entries of the miRNA dataset. We obtained miRNA expression data from TCGA, which is a collaboration between the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) that provides generic multidimensional maps of major genomic changes in 33 types of cancer [57]. To utilize the miRNA expression data for the weight matrix W, we first implemented min–max normalization for preprocessing. We entered a weight value into W only when the relationship between the miRNA i and disease j was zero. By mapping miRNA

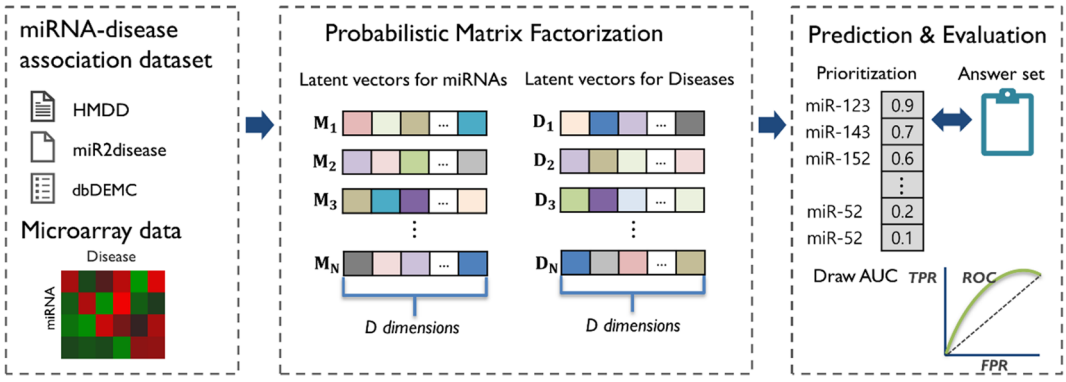

Fig. 1. Overall workflow of IMIPMF for disease-related miRNA prediction.

expression values onto objective functions, we can effectively learn latent vectors, although the corresponding known miRNA–disease associations do not exist.

2.2. Methods

2.2.1. Probabilistic matrix factorization

We adopted probabilistic matrix factorization (PMF) to handle limited miRNA data and improve the prediction accuracy. PMF is an algorithm based on matrix factorization that has already shown outstanding performance in recommender systems [58]. Recommender systems are widely used for selecting online information that is relevant to a specific user. Generally, in recommender systems, each user u rates a set of items based on their preferences. Considering the existing rating scores, a recommender system predicts a rating score for a user u for a non-rated item i to recommend new items that the user may like.

PMF is a probabilistic factor-based model for collaborative filtering that performs well on sparse and imbalanced datasets. In recommender systems, most datasets are composed of "infrequent" users who have rated fewer than five items and "frequent" users who have rated over 10,000 items. Collaborative filtering performs well when users have provided enough rating scores on several items, but it does not perform well for "cold start" users. Cold start users are new users with few ratings. The goal of predicting a rating score is analogous to predicting whether a particular miRNA is related to a specific disease or not. The cold start problem also exists in predicting disease-related miRNAs owing to the limited amount of data. Inspired by previous recommender system approaches that have handled the cold start problem, we apply PMF to overcome these obstacles.

The goal is to infer the most potential miRNA–disease associations for a particular disease. First, we define a conditional distribution over a set of given miRNA–disease associations as

$$p(R|M, D, \sigma^2) = \prod_{i=1}^{N_m} \prod_{j=1}^{N_d} [N(R_{ij}|M_i^T D_j, \sigma^2)]^{I_{ij}} \tag{1}$$

where $N(R_{ij}|M_i^T D_j, \sigma^2)$ is a probability density function that follows a Gaussian distribution with mean u and variance σ^2. We adopted zero-mean spherical Gaussian priors on miRNA and disease feature vectors as follows:

$$p(M|\sigma_M^2) = \prod_{i=1}^{N_m} N(M_i|0, \sigma_M^2 I) \tag{2}$$

$$p(D|\sigma_D^2) = \prod_{i=1}^{N_d} N(D_i|0, \sigma_D^2 I) \tag{3}$$

The log-likelihood of M and D are given by

$$\ln p(M, D|R, \sigma^2, \sigma_M^2, \sigma_D^2) = -\frac{1}{2\sigma^2} \sum_{i=1}^{N_m} \sum_{j=1}^{N_d} W_{ij} (R_{ij} - M_i^T D_j)^2 - \frac{1}{2\sigma_M^2} \sum_{i=1}^{N_m} M_i^T M_i - \frac{1}{2\sigma_D^2} \sum_{j=1}^{N_d} D_j^T D_j + Z \tag{4}$$

For the better prediction of miRNA–disease associations, we used the miRNA expression value for the weight of our cost function. W is the miRNA expression matrix whose entries are equal to the expression values for the existing miRNA–disease associations. For unknown associations, the matrix entries are zero. Here, Z is a constant term that does not depend on the parameters. Maximizing the log-posterior can be thought of as equivalent to minimizing the following objective function:

$$E = \frac{1}{2} \sum_{i=1}^{N_m} \sum_{j=1}^{N_d} W_{ij} (R_{ij} - M_i^T D_j)^2 + \frac{\lambda_M}{2} \sum_{i=1}^{N_m} ||M_i||^2 + \frac{\lambda_D}{2} \sum_{j=1}^{N_d} ||D_j||^2 \tag{5}$$

The first term is the squared error and the following two terms are regularization terms, where $\lambda_M = \sigma^2/\sigma_M^2$ and $\lambda_D = \sigma^2/\sigma_D^2$. Instead of using a simple linear-Gaussian model, we utilize the dot product of the miRNA and disease feature vectors that pass through the logistic function $g(x) = 1/[1 + \exp(-x)]$.

2.2.2. Constrained PMF

miRNAs with a few known disease associations will have feature vectors that are close to the prior mean. Therefore, we propose a novel method for handling miRNA-specific feature vectors for "infrequent" miRNAs. An infrequent miRNA is an miRNA that has few known associations with diseases. In general, such miRNAs lead to low prediction accuracy based on the limited number of entries in the dataset. These undesirable anomalies can be handled by adding a constrained model. Our constrained model reflects the conventional biological assumption that similar miRNAs are inclined to associate with phenotypically similar diseases.

We define the miRNA feature vector M_i as

$$M_i = Y_i + \frac{\sum_{k=1}^{N_d} I_{ik} C_k}{\sum_{k=1}^{N_d} I_{ik}} \tag{6}$$

where $C \in R^{N_l \times N_d}$ is a latent similarity constraint matrix with a zero-mean spherical-Gaussian prior defined as

$$p(C|\sigma_C^2) = \prod_{k=1}^{N_d} N(C_k|0, \sigma_w^2 I) \tag{7}$$

where I is an indicator matrix, in which $I_{ij} = 1$ if miRNA i has been found to be related to disease j. Otherwise, the entries are equal to zero. C is a latent variable that every miRNA shares. By adopting the constrained model, miRNAs with a few disease associations can obtain information from other miRNAs with abundant disease associations through the latent variable C. That is, the variable C works for sharing information among the miRNAs. Y_i can be regarded as an offset that is added to the mean of the prior distribution to derive a new M_i that has a strong effect on handling infrequent miRNAs. Without using this constraint, Y_i and M_i would be equal if the mean was fixed at zero. That is, Y_i would be the same as M_i without considering the constraint. The new M_i can be replaced with $g(Y_i) + \frac{\sum_{k=1}^{N_d} I_{ik} C_k}{\sum_{k=1}^{N_d} I_{ik}}$ by adopting the constrained model. We tuned our miRNA vector M_i more precisely with the new latent variable C to better capture the properties of miRNAs in a technical way. $g(x) = 1/[1 + \exp(-x)]$ is the logistic function, as mentioned above. The notations that were used in the equations are described in Table 1. A graphical representation of the constrained model is presented in Fig. 2. Now, the new constrained conditional distribution can be defined as follows:

$$p(R|Y, D, C, \sigma^2) = \prod_{i=1}^{N_m} \prod_{j=1}^{N_d} \left[N\left(R_{ij} | g\left(\left[Y_i + \frac{\sum_{k=1}^{N_d} I_{ik} C_k}{\sum_{k=1}^{N_d} I_{ik}}\right]^T D_j\right), \sigma^2\right) \right]^{I_{ij}} \tag{8}$$

3. Results

To evaluate the prediction accuracy of IMIPMF, we implemented 5-fold cross-validation. First, we divided the combined miRNA–disease association datasets into train and test sets with the ratio of 80/20.

Table 1
Notations.

Symbol	Description
N_m	Number of miRNAs
N_d	Number of diseases
N_l	Number of latent dimensions
$M \in R^{N_m \times N_l}$	miRNA latent space
$D \in R^{N_d \times N_l}$	Disease latent space
$C \in R^{N_d \times N_l}$	Latent similarity constraint matrix

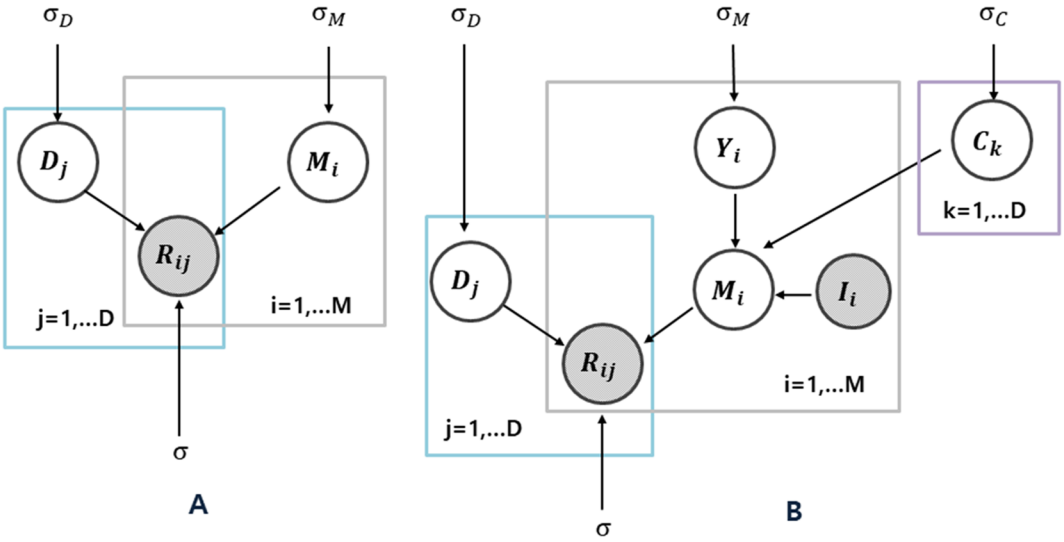

Fig. 2. Graphical model for (A) PMF and (B) PMF using constrained model.

Owing to the randomness in the choice of samples, cross-validation was implemented repeatedly to derive the mean "area under the receiver operating characteristic (ROC) curve" (AUC) scores. To analyze the performance intuitively, we drew the ROC curves by plotting the true positive rate (sensitivity) and false positive rate (1 − specificity), where sensitivity refers to the percentage of disease-related miRNAs and specificity refers to the percentage of miRNAs that were not found to be related to a particular disease. The TPR refers to the percentage of correctly identified true disease-related miRNAs and the FPR refers to the percentage of the disease-related miRNAs that are not identified so. The AUC values were calculated to compare the performance of each model. Generally, an AUC value of 1 refers to perfect prediction and 0.5 refers to the results of random choice. IMIPMF accomplished an AUC value of 0.891, which proves the superiority of IMIPMF compared to eight state-of-the-art methods.

3.1. Performance comparisons with other methods

To validate the predictive power of IMIPMF for verifying miRNA–disease associations, we compared its performance to those of eight state-of-the-art methods. We first calculated the AUC score for each model. As illustrated in Fig. 3(a), our IMIPMF model achieved superior performance compared to the IMCMDA [38], AMVML [44], PMAMCA [34], GRNMF [51], WBSMDA [26], RWRMDA [21], RLSMDA [25], and HDMP [23] models, with a reliable value of 0.891.

For further evaluation, we carried out leave-one-out cross-validation (LOOCV). The ROC curve based on LOOCV is illustrated in Fig. 3(b). After drawing the ROC curve, the area under the curve (AUC) was calculated for the comparison of IMIPMF with the eight state-of-the-art methods. LOOCV regards a single sample, in turn, as the test data, whereas all the other remaining miRNA–disease pairs as the training set. This process was repeated such that every sample was used as a validation sample at least once. In comparison with other previous prediction models on LOOCV, IMIPMF achieved the best AUC value of 0.911. Extensive experiments on the additional evaluation metric showed significant improvements in our IMIPMF framework over the state-of-the-art methods in terms of prediction accuracy and performance stability.

In contrast to the rapidly increasing number of newly discovered miRNAs, only a few miRNA–disease associations are known. Despite the various prominent properties of miRNAs in disease incidence, most computational models are highly dependent on known miRNA–disease associations. However, miRNAs with no previously known disease associations or a few disease associations might lead to low prediction accuracy. Such miRNAs are defined as infrequent miRNAs. To handle this issue, we developed our constrained model to control infrequent miRNAs. The performance of IMIPMF was improved by adding a new latent variable C (latent similarity constraint matrix) to the constrained model. As illustrated in Fig. 4, application of the PMF model alone resulted in an AUC value of 0.873. However, by adopting the constrained model, the AUC value was improved to 0.891. This is a remarkable result, considering that the best result was obtained using only two datasets (human miRNA–disease associations and miRNA expression data) with the constrained model.

3.2. Case studies

3.2.1. Breast cancer

Cancer develops when mutations that regulate cell growth occur in genes. According to the Center for Disease Control and Prevention, breast cancer is the most common female malignant neoplasm that compromises 30% of female cancer [59]. Based on its frequent occurrence, we prioritized the top-50 breast-cancer-related miRNAs to analyze the causes of breast cancer. Consequently, as shown in Table 2, IMIPMF extracted 49 true miRNA–disease associations according to our integrated answer set data. Moreover literature-based analysis was implemented to determine whether the remaining candidate had any potential involvement in the incidence of breast cancer. Surprisingly, miR-142 (miR-142-3p) has exhibited a dysregulated presentation in various subtypes of breast cancer. Various studies have also reported that an overexpression of miR-142 might result in the down-regulation of the genes WASL and RAC1, which are known to be related to cell mortality [60]. Furthermore, miR-142 was also confirmed to play a crucial role in inhibiting the invasiveness of breast cancer cells. Based on these findings, we were able to confirm that our top 50 candidates are all related to breast cancer.

A significant criterion for estimating the capability of the model is whether or not IMIPMF can be applied to handle miRNAs with no disease associations. For the precise evaluation of miRNAs without any known diseases, we picked several miRNAs that were proved to be true

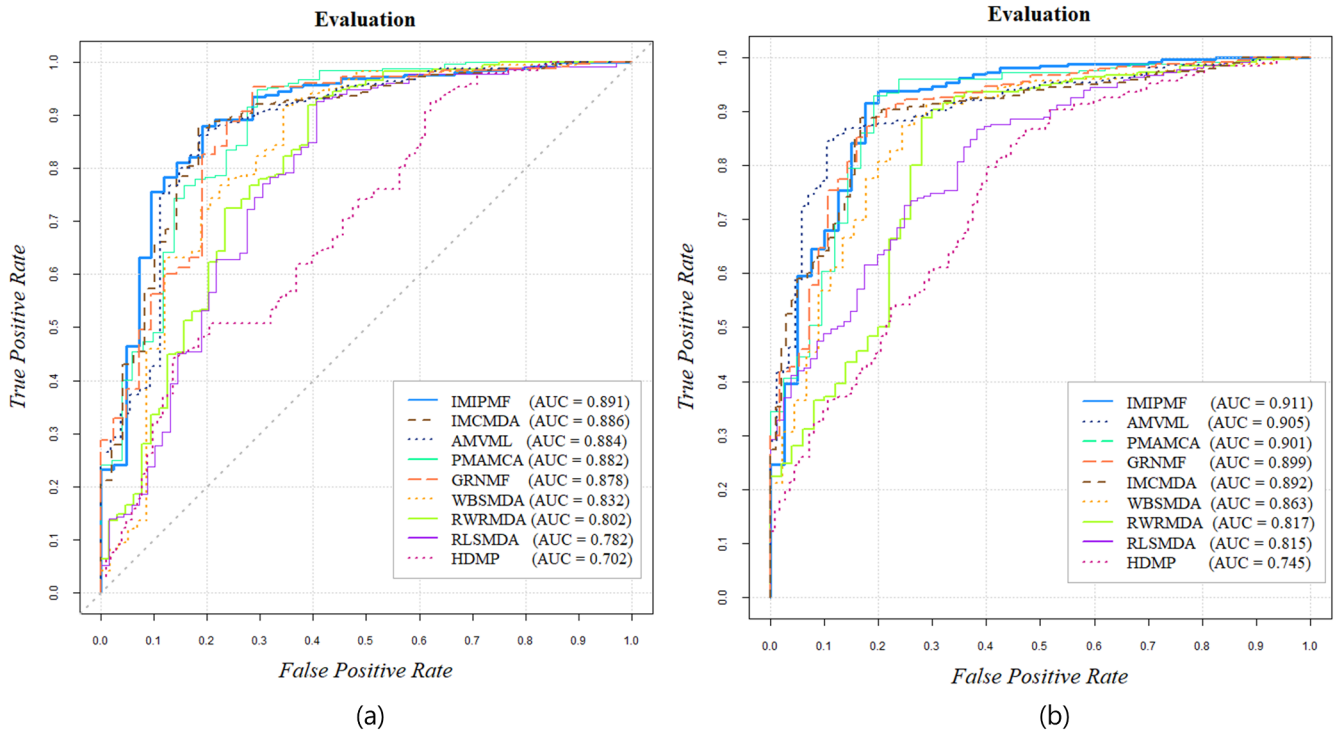

Fig. 3. Performance comparison between IMIPMF and eight state-of-the-art miRNA–disease association prediction models in terms of ROC curves and AUC values in (a)5-fold cross-validation and (b)LOOCV.

disease-related miRNAs by IMIPMF and manually removed their labels to produce a miRNA with no disease associations. Thereafter, we performed IMIPMF to check whether or not our method predicts the relationship between the new miRNAs and specific diseases well. For example, we removed the labels on hsa-miR-155 and hsa-let-7d, which were high-ranked breast-cancer-related miRNAs in Table 2. Surprisingly, although we removed the labels on hsa-miR-155 and hsa-let-7d, these miRNAs still proved to be breast-cancer-related miRNAs by IMIPMF. The success of IMIPMF in predicting new miRNAs can be explained by its mathematical foundations in MF and the utilization of miRNA expression data.

3.2.2. Lung cancer

Lung cancer is the leading cause of death worldwide despite numerous advances in surgical treatment [61]. There is an increasing requirement for detecting biomarkers that can help researchers understand the biological mechanisms of lung cancer. Therefore, IMIPMF was used to prioritize the top 50 candidate miRNAs that have a high chance of being involved in lung cancer incidence. Surprisingly, among our top 50 candidates, 49 miRNAs were found to be true lung-cancer-related miRNAs based on our combined answer set data. A list of the confirmed miRNAs is provided in Table 3. We also performed literature-based analysis to determine if the remaining miRNAs had any potential involvement in lung cancer incidence. According to published experimental studies, miR-127 may play an essential role in lung adenocarcinoma and poor prognoses [62]. Furthermore, high levels of miR-127 might lead to stem-like transitions, indicating that miR-127 is related to the formation of the aggressive phenotypes of lung cancer.

We performed functional enrichment analysis on miR-127 using an online enrichment tool (TAM). TAM is a web-based miRNA functional enrichment tool that returns the biological meaning and common functions of a specific query miRNA [63]. Based on this enrichment test, we found that miR-127 plays a significant role in the occurrence of breast cancer. Generally, lung cancer is known to be a disease phenotypically similar to breast cancer. These results also demonstrate the biological assumption that functionally related miRNAs are inclined to associate with phenotypically similar diseases. Furthermore, we also demonstrated our performance by prioritizing the top 50 colon-cancer-related candidates. Among the 50 candidates, 46 candidates were validated to be true disease-related miRNAs based on dbDEMC. The table providing the result is included in the Supplementary Material. By considering these results, we validated the excellent performance of IMIPMF for detecting disease-related miRNAs and inferring potential miRNA–disease associations.

Further, the performance of the new miRNAs was tested. We manually removed the labels on hsa-miR-146a and hsa-miR-133b, which were high-ranked lung-cancer-related miRNAs. After removing the labels, we implemented IMIPMF to check its performance on new miRNAs. Surprisingly, these two miRNAs were also proved to be lung-cancer-related miRNAs by IMIPMF. These experiments address the aforementioned research problems by taking the inner product of the latent vectors, which indirectly reflect the potential miRNA–disease associations.

3.2.3. Kaplan–Meier survival analysis

Identification of the association between miRNAs and the prognosis of breast cancer patients is a vital process in understanding disease pathogenesis [64,65,66]. To perform the Kaplan–Meier survival analysis, we utilized the miRpower-Kaplan–Meier plotter web-tool [67]. For precise analysis, we selected the TCGA dataset and only those miRNAs with a p-value < 0.005 were considered as highly associated with the overall survival of breast cancer patients. The Kaplan–Meier survival analysis of the miRNA candidates predicted by IMIPMF proved that the high-ranked miRNAs, hsa-miR-148a, hsa-miR-133b, hsa-miR-

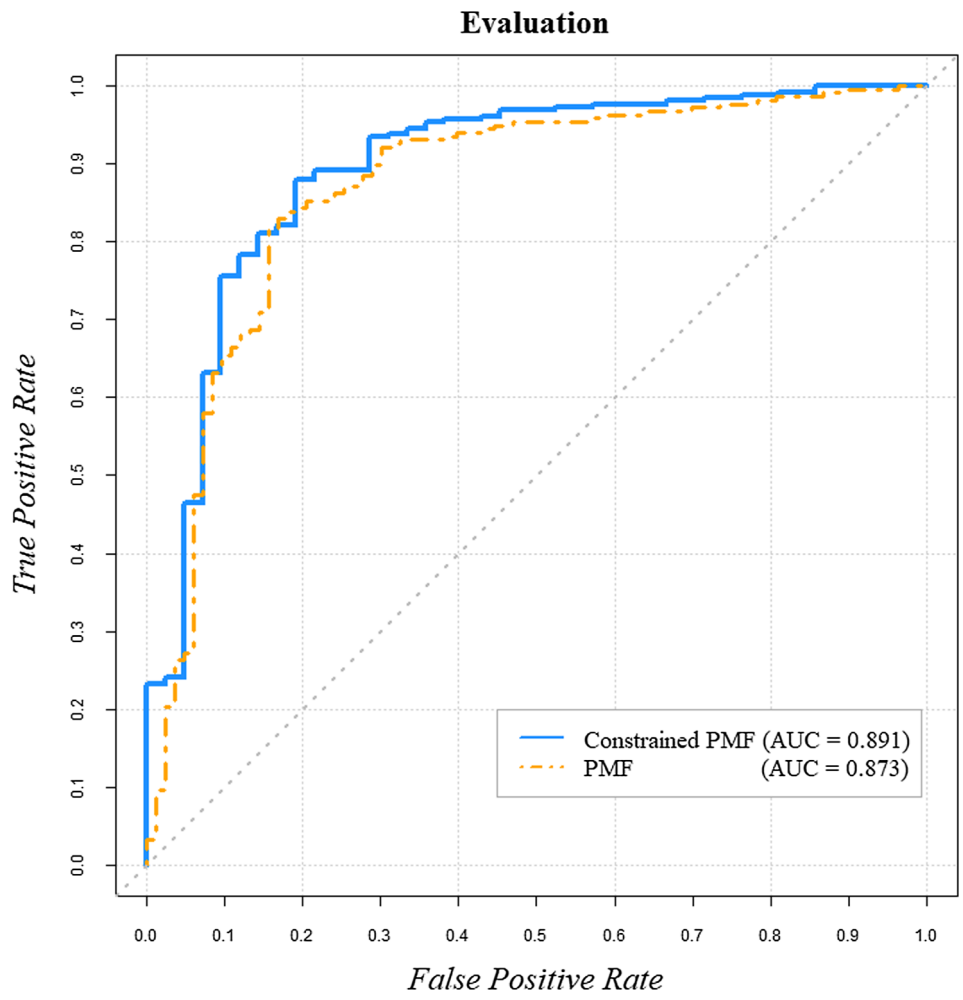

Fig. 4. Performance comparison between original PMF and PMF with constrained model. Application of constrained model increased AUC score.

Table 2
Top 50 breast-cancer-related miRNAs confirmed by IMIPMF. 5-fold cross-validation was implemented using public databases (dbDEMC, HMDD, and miR2Disease) and literature-based analysis. All 50 miRNAs were proven to be associated with breast cancer.

Rank	Name	Evidence	Rank	Name	Evidence
1	hsa-miR-146a	miR2Disease, dbDEMC	26	hsa-miR-139	dbDEMC
2	hsa-miR-155	miR2Disease, dbDEMC	27	hsa-miR-130a	dbDEMC
3	hsa-miR-16	dbDEMC	28	hsa-let-7f	miR2Disease, dbDEMC
4	hsa-miR-148b	dbDEMC	29	hsa-let-7i	miR2Disease, dbDEMC
5	hsa-miR-181a	miR2Disease, dbDEMC	30	hsa-miR-142	Literature [60]
6	hsa-miR-145	miR2Disease, dbDEMC	31	hsa-miR-153	dbDEMC
7	hsa-let-7g	dbDEMC	32	hsa-miR-106a	dbDEMC
8	hsa-miR-151	dbDEMC	33	hsa-miR-191	miR2Disease, dbDEMC
9	hsa-miR-126	miR2Disease, dbDEMC	34	hsa-miR-1274a	dbDEMC
10	hsa-miR-125a	miR2Disease, dbDEMC	35	hsa-miR-1181	dbDEMC
11	hsa-miR-18a	miR2Disease, dbDEMC	36	hsa-miR-134	dbDEMC
12	hsa-let-7e	dbDEMC	37	hsa-miR-17	dbDEMC
13	hsa-let-7a	miR2Disease, dbDEMC	38	hsa-miR-10a	dbDEMC
14	hsa-miR-181b	miR2Disease, dbDEMC	39	hsa-miR-135a	dbDEMC
15	hsa-miR-10b	miR2Disease, dbDEMC	40	hsa-miR-1226*	dbDEMC
16	hsa-let-7d	miR2Disease, dbDEMC	41	hsa-miR-107	dbDEMC
17	hsa-miR-183	dbDEMC	42	hsa-miR-187*	dbDEMC
18	hsa-miR-18b	dbDEMC	43	hsa-miR-1182	dbDEMC
19	hsa-miR-1247	dbDEMC	44	hsa-miR-184	dbDEMC
20	hsa-miR-125b	miR2Disease, dbDEMC	45	hsa-miR-149	miR2Disease, dbDEMC
21	hsa-miR-15a	dbDEMC	46	hsa-miR-135b	dbDEMC
22	hsa-miR-1303	dbDEMC	47	hsa-miR-185	dbDEMC
23	hsa-miR-1282	dbDEMC	48	hsa-miR-150	dbDEMC
24	hsa-miR-155*	dbDEMC	49	hsa-miR-100	dbDEMC
25	hsa-miR-127	miR2Disease, dbDEMC	50	hsa-miR-1206	dbDEMC

Table 3
Top 50 candidate lung-cancer-related miRNAs confirmed by IMIPMF. Validation was performed using various public databases and literature-based analysis. All 50 candidates were found to be lung-cancer-related miRNAs.

Rank	Name	Evidence	Rank	Name	Evidence
1	hsa-miR-155	miR2Disease, dbDEMC	26	hsa-miR-126	miR2Disease, dbDEMC
2	hsa-miR-148a	dbDEMC	27	hsa-miR-18b	dbDEMC
3	hsa-miR-100	dbDEMC	28	hsa-let-7i	dbDEMC
4	hsa-miR-149	dbDEMC	29	hsa-miR-188	dbDEMC
5	hsa-miR-146a	miR2Disease, dbDEMC	30	hsa-miR-181b	dbDEMC
6	hsa-miR-146b	miR2Disease, dbDEMC	31	hsa-let-7c	miR2Disease, dbDEMC
7	hsa-miR-17	dbDEMC	32	hsa-miR-127	Literature [62]
8	hsa-miR-187	dbDEMC	33	hsa-miR-181a	dbDEMC
9	hsa-miR-133b	miR2Disease, dbDEMC	34	hsa-let-7g	miR2Disease, dbDEMC
10	hsa-miR-125b	dbDEMC	35	hsa-miR-101	miR2Disease, dbDEMC
11	hsa-miR-125a	miR2Disease, dbDEMC	36	hsa-miR-181c	miR2Disease, dbDEMC
12	hsa-miR-107	dbDEMC	37	hsa-miR-186	dbDEMC
13	hsa-miR-152	dbDEMC	38	hsa-miR-129	dbDEMC
14	hsa-miR-16	miR2Disease, dbDEMC	39	hsa-miR-144	dbDEMC
15	hsa-miR-132	dbDEMC	40	hsa-miR-1234	dbDEMC
16	hsa-miR-1	miR2Disease, dbDEMC	41	hsa-miR-184	dbDEMC
17	hsa-miR-130a	miR2Disease, dbDEMC	42	hsa-miR-128	dbDEMC
18	hsa-miR-130b	dbDEMC	43	hsa-miR-145	miR2Disease, dbDEMC
19	hsa-let-7b	miR2Disease, dbDEMC	44	hsa-miR-1247	dbDEMC
20	hsa-let-7e	miR2Disease, dbDEMC	45	hsa-miR-1301	dbDEMC
21	hsa-miR-15a	dbDEMC	46	hsa-miR-106b	dbDEMC
22	hsa-miR-150	miR2Disease, dbDEMC	47	hsa-miR-182*	miR2Disease, dbDEMC
23	hsa-let-7a	miR2Disease, dbDEMC	48	hsa-miR-124	dbDEMC
24	hsa-miR-15b	dbDEMC	49	hsa-miR-10a	dbDEMC
25	hsa-miR-185	dbDEMC	50	hsa-miR-1292	dbDEMC

125b, and hsa-miR-125a, were considerably related to the overall survival of breast cancer patients (see Fig. 5).

3.2.4. *t*-distributed stochastic neighbor embedding visualization

To investigate whether the latent vectors play a significant role in classifying disease categories, we visualized the latent vectors learned by IMIPMF using *t*-distributed stochastic neighbor embedding (*t*-SNE) [68]. Visualization of the latent vectors revealed the characteristics of miRNAs for different disease categories and expressed the underlying within-disease similarities. In Fig. 6, each point represents a latent vector and each color represents a disease category. It can be seen that points in the same disease category tend to group together based on the correlation of the disease categories.

More significantly, we were able to confirm that phenotypically similar diseases are inclined to be located near each other. Lung cancer and breast cancer are known as phenotypically similar diseases, as confirmed by MimMiner [69]. This supports the famous biological assumption that functionally related miRNAs are inclined to associate with phenotypically similar diseases. Overall, this experiment verified that the proposed model not only predicts the relationships between diseases and miRNAs but also has the potential to calculate disease similarities.

4. Discussion

There is significant evidence has to show that miRNAs play pivotal roles in regulating key biological functions as well as disease incidence. Several computational methods have been proposed to search for novel miRNA–disease associations and enhance prediction accuracy. This paper presents a novel computational model termed IMIPMF that uses the machine learning technique PMF to assign scores to miRNA–disease pairs. Through the application of PMF and miRNA expression data, we effectively inferred miRNA–disease associations, even without base knowledge about miRNA–disease associations. Furthermore, by assuming that similar miRNAs are inclined to associate with phenotypically similar diseases, we adopted a constrained model to reduce the effects of imbalanced datasets. The performance of the IMIPMF method was demonstrated by implementing 5-fold cross-validation. To compare the performance of our model with that of previous state-of-the-art methods intuitively, we calculated the AUC scores by drawing ROC curves. In addition, experiments were performed on various critical human diseases, namely breast cancer, lung cancer, brain cancer, colon cancer, and kidney cancer. The top 50 and top 10 candidate miRNAs were identified based on the HMDD, dbDEMC, and miR2Disease databases and literature-based analysis. Therefore, IMIPMF can be used as a biological tool for extracting novel disease-related miRNAs, thereby providing biological insights into the disease mechanisms of miRNAs and facilitating the early diagnosis and treatment of human diseases in the future.

5. Conclusions

The effective performance of IMIPMF depends on several factors. First, as meaningful biological data continue to be generated, our model can apply new miRNA expression data as objective function weights to efficiently learn latent vectors, even if we do not have any information on previously known miRNA–disease associations. That is, our model is not entirely dependent on known disease-related miRNA samples. Second, using a constrained model, we mitigate the impact of infrequent miRNAs to improve the prediction accuracy of IMIPMF. Most significantly, our model utilizes a machine learning technique termed PMF that has achieved excellent performance in recommender systems. Most companies have gained vast profits by adopting recommender systems. The adoption of PMF not only improves prediction accuracy but also facilitates the identification of novel disease-related miRNA candidates. There is still room for enhancing the prediction power of the proposed model by adding implicit feedback such as information regarding target-gene and RNA-sequence data. Further, extracting

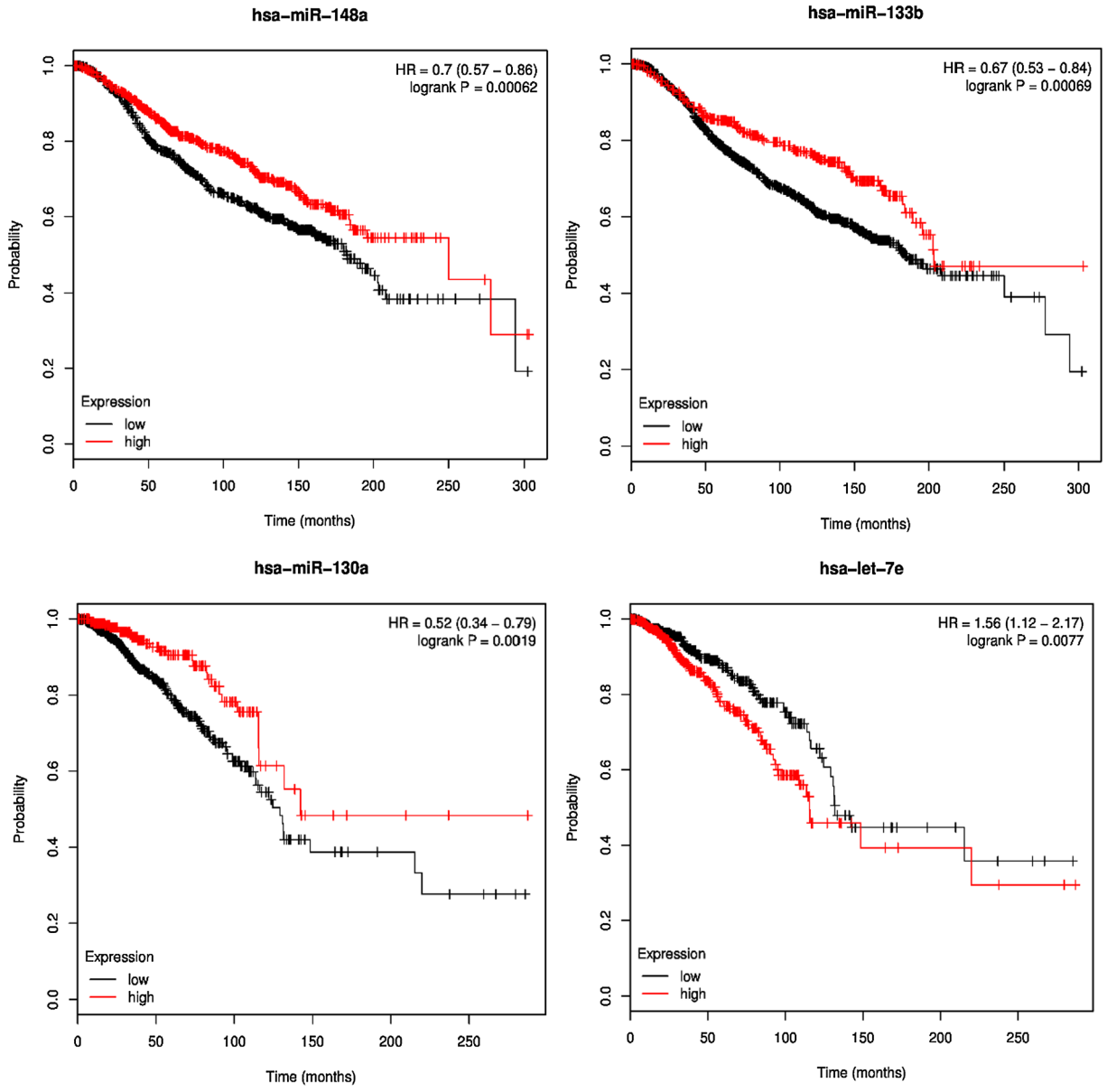

Fig. 5. Kaplan–Meier plots of hsa-miR-148a, hsa-miR-133b, hsa-miR-130a, and hsa-let-7e for survival of breast cancer patients.

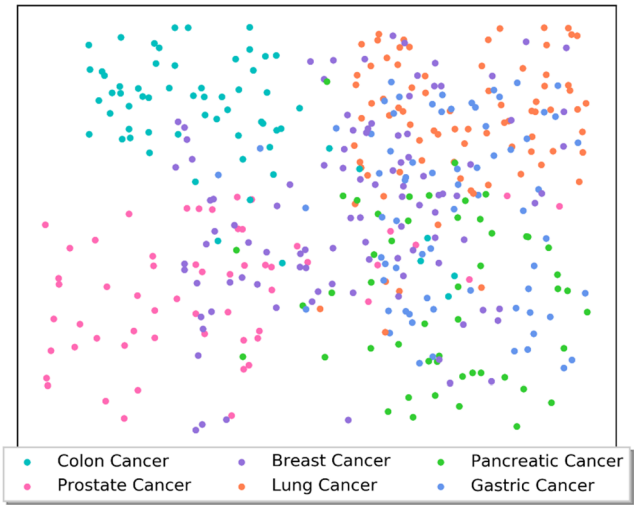

Fig. 6. Visualization of latent vectors learnt by IMIPMF.

meaningful features from various biological data could lead to better performance in the future.

CRediT authorship contribution statement

Jihwan Ha: Conceptualization, Data curation, Methodology, Validation, Visualization, Writing - original draft. **Chihyun Park:** Methodology, Investigation, Writing - review & editing. **Chanyoung Park:** Formal analysis, Investigation, Writing - review & editing. **Sanghyun Park:** Supervision, Writing - review & editing, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

Funding: This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the SW Starlab support program [IITP-2017-0-00477] supervised by the IITP (Institute for Information & Communications Technology Promotion).

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jbi.2019.103358.

References

[1] V. Ambros, The functions of animal microRNAs, Nature 431 (2004) 350–355.
[2] D.P. Bartel, MicroRNAs: genomics, biogenesis, mechanism, and function, Cell 116 (2004) 281–297.
[3] M. Alshalalfa, R. Alhajj, Using context-specific effect of miRNAs to identify functional associations between miRNAs and gene signatures, BMC Bioinf. 14 (2013) S1.
[4] D.P. Bartel, MicroRNAs: target recognition and regulatory functions, Cell 136 (2009) 215–233.
[5] P. Xu, M. Guo, B.A. Hay, MicroRNAs and the regulation of cell death, Trends Genet 20 (2004) 617–624.
[6] X. Karp, V. Ambros, Encountering microRNAs in cell fate signaling, Science 310 (2005) 1288–1289.
[7] E.A. Miska, How microRNAs control cell division, differentiation and death, Curr. Opin. Genet. Dev. 15 (2005) 563–568.
[8] A.M. Cheng, M.W. Byrom, J. Shelton, L.P. Ford, Antisense inhibition of human miRNAs and indications for an involvement of miRNA in cell growth and apoptosis, Nucleic Acids Res. 33 (2005) 1290–1297.
[9] Q. Cui, Z. Yu, E.O. Purisima, E. Wang, Principles of microRNA regulation of a human cellular signaling network, Mol. Syst. Biol. 2 (2006) 46.
[10] S. Griffiths-Jones, miRBase: microRNA sequences and annotation, Curr. Protoc. Bioinform. 12 (2010) 1–10.
[11] R. Wang, et al., MiR-101 is involved in human breast carcinogenesis by targeting Stathmin1, PLoS ONE 7 (2012) e46173.
[12] Y. Akao, Y. Nakagawa, et al., Role of anti-oncomirs miR-143 and -145 in human colorectal tumors, Cancer Gene Ther. 17 (2010) 398–408.
[13] R. Wang, et al., MiR-185 is involved in human breast carcinogenesis by targeting Vegfa, FEBS Lett. 588 (2014) 4438–4447.
[14] K.J. Png, et al., MicroRNA-335 inhibits tumor reinitiation and is silenced through genetic and epigenetic mechanisms in human breast cancer, Genes Dev. 25 (2011) 226–231.
[15] S.F. Tavazoie, et al., Endogenous human microRNAs that suppress breast cancer metastasis, Nature 451 (2008) 147–152.
[16] S. Valastyan, et al., A pleiotropically acting microRNA, miR-31, inhibits breast cancer metastasis, Cell 137 (2009) 1032–1046.
[17] C. Perez-Iratxeta, M. Wjst, P. Bork, M.A. Andrade, G2D: a tool for mining genes associated with disease, BMC Gene 6 (2005) 45.
[18] C. Perez-Iratxeta, P. Bork, M.A. Andrade, Association of genes to genetically inherited diseases using data mining, Nat. Gene 31 (2002) 316–319.
[19] S. Aerts, et al., Gene prioritization through genomic data fusion, Nat. Biotechnol. 24 (2006) 537–544.
[20] Q. Jiang, et al., Prioritization of disease microRNAs through a human phenome-microRNAome network, BMC Syst. Biol. 4 (2010) S2.
[21] X. Chen, M.-X. Liu, G.-Y. Yan, RWRMDA: predicting novel human microRNA–disease associations, Mol. Biosyst. 8 (2012) 2792–2798.
[22] S. Mørk, S. Pletscher-Frankild, A.P. Caro, J. Gorodkin, L.J. Jensen, Protein-driven inference of miRNA–disease associations, Bioinformatics 30 (2013) 392–397.
[23] P. Xuan, et al., Prediction of microRNAs associated with human diseases based on weighted k most similar neighbors, PLoS ONE 8 (2013) e70204.
[24] X. Chen, C. Yan, X. Zhang, Z. You, Y. Huang, G. Yan, HGIMDA: heterogeneous graph inference for MiRNA-disease association prediction, Oncotarget 7 (2016) 65257–65269, https://doi.org/10.18632/oncotarget.11251.
[25] X. Chen, G.-Y. Yan, Semi-supervised learning for potential human microRNA-disease associations inference, Sci. Rep. 4 (2014) 5501.
[26] X. Chen, C.C. Yan, X. Zhang, Z.H. You, L. Deng, Y. Liu, Y. Zhang, Q. Dai, WBSMDA: within and between score for MiRNA-disease association prediction, Sci. Rep. 6 (2016) 21106.
[27] G. Li, et al., Predicting microRNA-disease associations using label propagation based on linear neighborhood similarity, J. Biomed. Informat. 82 (2018) 169–177.
[28] M. Aqil, A.R. Naqvi, S. Mallik, S. Bandyopadhyay, U. Maulik, S. Jameel, The HIV Nef protein modulates cellular and exosomal miRNA profiles in human monocytic cells, J. Extracell Vesicles (2014) 3.
[29] M. Aqil, S. Mallik, S. Bandyopadhyay, U. Maulik, S. Jameel, Transcriptomic analysis of mRNAs in human monocytic cells expressing the HIV-1 Nef protein and their exosomes, Biomed. Res. Int. 2015 (2015) 492395.
[30] X. Shu, L. Cheng, Z. Dong, S. Shu, Identification of circular RNA-associated competing endogenous RNA network in the development of cleft palate, J. Cell Biochem. 120 (2019) 16062–16074.
[31] S. Sen, U. Maulik, S. Mallik, S. Bandyopadhyay, Detecting TF-MiRNA-gene network based modules for 5hmC and 5mC brain samples: a intra- and inter-species case-study between human and rhesus, BMC Genet. (2017), https://doi.org/10.1186/s12863-017-0574-7.
[32] N.J. Martinez, A.J. Walhout, The interplay between transcription factors and microRNAs in genome-scale regulatory networks, BioEssays 31 (2009) 435–445.
[33] J. Ha, H. Kim, Y. Yoon, S. Park, A method of extracting disease-related microRNAs through the propagation algorithm using the environmental factor based global miRNA network, Bio-Med. Mater. Eng. 26 (s1) (2015) S1763–S1772.
[34] J. Ha, C. Park, S. Park, PMAMCA: prediction of microRNA-disease association utilizing a matrix completion approach, BMC Syst. Biol. 13 (1) (2019).
[35] X. Chen, et al., RBMMMDA: predicting multiple types of disease-microRNA associations, Sci. Rep. 5 (2015) 13877.
[36] J.Q. Li, Z.H. Rong, X. Chen, G.Y. Yan, Z.H. You, MCMDA: matrix completion for MiRNA-disease association prediction, Oncotarget 8 (2017) 21187–21199.
[37] X. Chen, Y.W. Niu, G.H. Wang, G.Y. Yan, HAMDA: hybrid approach for MiRNA-disease association prediction, J. Biomed. Inform. 76 (2017) 50–58, https://doi.org/10.1016/j.jbi.2017.10.014.
[38] X. Chen, L. Wang, J. Qu, N.N. Guan, J.Q. Li, Predicting miRNA-disease association based on inductive matrix completion, Bioinformatics 34 (24) (2018) 4256–4265.
[39] X. Chen, J. Yin, J. Qu, L. Huang, MDHGI: matrix decomposition and heterogeneous graph inference for miRNA-disease association prediction, PLoS Comput. Biol. 14 (2018) e1006418, https://doi.org/10.1371/journal.pcbi.1006418 PMID: 30142158.
[40] X. Chen, L. Huang, LRSSLMDA: Laplacian regularized sparse subspace learning for MiRNA disease association prediction, PLoS Comput. Biol. 13 (2017) e1005912, https://doi.org/10.1371/journal.pcbi.1005912 PMID: 29253885.
[41] X. Chen, D. Xie, L. Wang, Q. Zhao, Z.H. You, et al., BNPMDA: bipartite network projection for MiRNA-disease association prediction, Bioinformatics (Oxford, England) 34 (2018) 3178–3186, https://doi.org/10.1093/bioinformatics/bty333 PMID: 29701758.
[42] X. Chen, D. Xie, Q. Zhao, Z.H. You, MicroRNAs and complex diseases: from experimental results to computational models, Briefings Bioinf. 20 (2019) 515–539, https://doi.org/10.1093/bib/bbx130 PMID: 29045685.
[43] X. Chen, L. Huang, D. Xie, Q. Zhao, EGBMMDA: extreme gradient boosting machine for MiRNA disease association prediction, Cell Death Dis. 9 (2018) 3, https://doi.org/10.1038/s41419-017-0003-x PMID: 29305594.
[44] C. Liang, S. Yu, J. Luo, Adaptive multi-view multi-label learning for identifying disease-associated candidate miRNAs, PLoS Comput Biol. 15 (4) (2019) e1006931.
[45] S. Shamsizadeh, S. Goliaei, ZahraRazaghi Moghadam, CAMIRADA: cancer microRNA association discovery algorithm, a case study on breast cancer, J. Biomed. Inform. 94 (2019) 103180.
[46] S. Mallik, Z. Zhao, Graph- and rule-based learning algorithms: a comprehensive review of their applications for cancer type classification and prognosis using genomic data, Brief. Bioinform. (2018) bby120.
[47] W. Liu, T. Wang, S. Chen, A. Tang, Feature Extraction and Discovery of microRNAs Using Nonnegative Matrix Factorization, Proceeding of the 11th joint conference on information sciences, (2018).
[48] X. Chen, S. Li, J. Yin, C. Wang, Potential miRNA-disease association prediction based on kernelized Bayesian matrix factorization, Genomics (2019), https://doi.org/10.1016/j.ygeno.2019.05.021.
[49] M. Gao, Z. Cui, Y. Gao, J. Liu, C. Zheng, Dual-network sparse graph regularized matrix factorization for predicting miRNA-disease associations, Mol. Omics 15 (2019) 130–137.
[50] Y. Gao, Z. Cui, J. Liu, J. Wang, C. Zheng, NPCMF: nearest profile-based collaborative matrix factorization method for predicting miRNA-disease associations, BMC Bioinf. 20 (2019) 353.
[51] Q. Xiao, J. Luo, C. Liang, J. Cai, P. Ding, A graph regularized non-negative matrix factorization method for identifying microRNA-disease associations, Bioinformatics (Oxford, England) (2017), https://doi.org/10.1093/bioinformatics/btx545 PMID: 28968779.
[52] M. Gönen, S. Kaski, Kernelized bayesian matrix factorization, IEEE Trans. Patt. Anal. Mach. Intell. 36 (10) (2014) 2047–2060.
[53] T. Zhou, H. Shan, A. Banerjee, G. Sapiro, Kernelized probabilistic matrix factorization: exploiting graphs and side information, Proceedings of SDM, 2012, pp. 403–414.
[54] Z. Yang, F. Ren, C. Liu, S. He, G. Sun, Q. Gao, L. Yao, Y. Zhang, R. Miao, Y. Cao, et al., dbDEMC: a database of differentially expressed miRNAs in human cancers, BMC Geno. 11 (Suppl 4) (2010) S5.
[55] Y. Li, C. Qiu, J. Tu, B. Geng, J. Yang, T. Jiang, Q. Cui, HMDD v2.0: a database for experimentally supported human microRNA and disease associations, Nucleic Acids Res. 42 (2014) (Database issue):D1070–4.
[56] Q. Jiang, Y. Wang, Y. Hao, L. Juan, M. Teng, X. Zhang, M. Li, G. Wang, Y. Liu, miR2Disease: a manually curated database for microRNA deregulation in human disease, Nucleic Acids Res. 37 (Database) (2009) D98–D104.
[57] K. Tomczak, P. Czerwinska, M. Wiznerowicz, The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge, Wspolczesna Onkol. 1A (2015) A68–A77, https://doi.org/10.5114/wo.2014.47136.
[58] Salakhutdinov, R.; Mnih, A. Probabilistic matrix factorization. In Advances in Neural Information Processing Systems; NIPS, 2007; Vol. 20, pp 1257 − 1264.
[59] R.L. Siegel, K.D. Miller, A. Jemal, Cancer statistics, 2019, CA Cancer J. Clin. 69 (2019) 7.
[60] A. Schwickert, et al., microRNA miR-142-3p inhibits breast cancer cell invasiveness by synchronous targeting of WASL, integrin alpha V, and additional cytoskeletal elements, PLoS One 10 (12) (2015).
[61] W.D. Travis, L.B. Travis, S.S. Devesa, Lung cancer [published erratum appears in Cancer 1995;75:2979], Cancer 75 (1 Suppl) (1995) 191–202.

[62] L. Shi, et al., miR-127 promotes EMT and stem-like traits in lung cancer through a feed-forward regulatory loop, Oncogene 36 (2017).
[63] M. Lu, B. Shi, J. Wang, Q. Cao, Q. Cui, TAM: a method for enrichment and depletion analysis of a microRNA category in a list of microRNAs, BMC Bioinf. 11 (2010) 419.
[64] X.H. Shi, X. Li, H. Zhang, et al., A five-microRNA signature for survival prognosis in pancreatic adenocarcinoma based on TCGA data, Sci. Rep. 8 (1) (2018) 7638.
[65] S. Bandyopadhyay, et al., A survey and comparative study of statistical tests for identifying differential expression from microarray data, IEEE/ACM Trans. Comput. Biol. Bioinformat. 11 (1) (2014) 95–115.
[66] S. Yerukala Sathipati, S.-Y. Ho, Identifying a miRNA signature for predicting the stage of breast cancer, Sci. Rep. 8 (1) (2018) 16138.
[67] A. Lanczky, et al., miRpower: a web-tool to validate survival-associated miRNAs utilizing expression data from 2178 breast cancer patients, Breast Cancer Res. Treat. 160 (2016) 439–446, https://doi.org/10.1007/s10549-016-4013-7.
[68] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," JMLR, 2008.
[69] M.A. Van Driel, J. Bruggeman, G. Vriend, et al., A text-mining analysis of the human phenome, Eur. J. Hum. Genet. 14 (2006) 535–542.