Pattern Recognition Letters 000 (2017) 1–8

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

DPFMDA: Distributed and privatized framework for miRNA-Disease association prediction

Lixin Chen, Bingtao Liu*, Chenggang Yan

Hangzhou Dianzi University, Hangzhou, China

ARTICLE INFO

Article history:
Available online xxx

MSC:
41A05
41A10
65D05
65D17

Keywords:
miRNA-disease association prediction
Computational model
Distributed data processing

ABSTRACT

In this paper, we developed a novel distributed and privatized framework for miRNA-disease association prediction (DPFMDA) to predict potential miRNA-disease associations. DPFMDA assumes that diseases and miRNAs can be represented as feature matrices. Through independent factorization, collaborative analysis and matrix re-factorization, feature matrices are estimated and potential associations in miRNA-disease association matrix are predicted. The proposed computational model is tested and verified by leave-one-out cross validation (LOOCV). The AUCs of global and local LOOCV are 0.8859 and 0.8267 respectively, which outperform the state-of-the-art computational models. DPFMDA is further evaluated with case studies of four important human complex diseases. We test Esophageal Neoplasms and Prostate Neoplasms using DPFMDA in the HMDD v2.0 database. For Breast Neoplasms, we use the HMDD v2.0 database which removes all the known miRNAs-disease associations on Breast Neoplasms. For Hepatocellular Carcinoma, HMDD v1.0 database is used as the input of DPFMDA and predicted results are verified with HMDD v2.0, dbDEMC and miR2Disease databases. The results show that 90% (Esophageal Neoplasms), 88% (Prostate Neoplasms), 98% (Breast Neoplasms) and 86% (Hepatocellular Carcinoma) of top 50 predicted miRNA-disease associations are confirmed by recent experimental studies respectively. It is anticipated that DPFMDA would be an effective and extensible method for potential miRNA-disease association prediction.

1. Introduction

MicroRNAs (miRNAs) are a class of small regulatory non-coding RNAs (~22nt) that have an impact on the expression of complementary mRNAs by binding to the 3 untranslated regions (UTRs) at post-transcriptional or translational level [15]. In addition, miRNAs could repress gene expression as well as function as positive regulators in some cases [27].

Accumulating studies indicated that miRNA plays a critical role in many biological processes, including the proliferation [4], development, differentiation and apoptosis [22] of cells, metabolism, aging, signal transduction, viral infection and so on. Therefore, the dysregulation of miRNAs is closely related to various human complex diseases [13], such as various cancers [14]. For example, Over-expression of miR-22 down-regulates ER and increases IL-1 expression, which promotes HBV-related hepatocellular carcinoma (HCC) development in males. MiR-122, as a liver-specific miRNA, is expressed highly in normal liver cell lines, but down-regulated in HCC derived cell lines, which makes it a promising biomarker for HCC diagnosis [18]. MiR-150 specifically targets the3'-UTR of p53 and regulates its expression, which promotes the proliferation of lung cancer cells [26]. Consequently, identifying disease-related miRNAs could benefit the diagnosis and prevention of human complex diseases, which is essential to biomedical research. However, searching and validating the possibility of miRNA-disease associations with biological experiments is costly and time-consuming. So it is necessary and efficient to uncover potential miRNA-disease assocaitions by diverse computational models.

In the last few years, with the updates to datasets by biological experiments and the development of various computational models based on the assumption that functionally similar miRNAs tend to be related to phenotypically similar diseases, the predictive accuracy increases [17]. Jiang et al. [8] developed a hypergeometric distribution based miRNA scoring system. Potential miRNA-disease associations can be identified after prioritizing miRNAs according to their score ranks. This computational model heavily relies on the predicted interactions between miRNAs and diseases, which are not reliable enough as the high rate of existing false-positive

* Corresponding author.
E-mail addresses: liubingtao@hdu.edu.cn (B. Liu), cgyan@hdu.edu.cn (C. Yan).

http://dx.doi.org/10.1016/j.patrec.2017.07.008
0167-8655/

and false-negative results. Shi et al. concerntrated on the functional links between disease genes and miRNA targets in protein-protein interaction (PPI) network and proposed a computational model based on a random walk algorithm. MiRNA-disease associations are systematically analyzed after constructing a bipartite miRNA-disease network. Mork et al. [16] analyzed miRNA-Protein-Disease associations in their miPRD computational model. MiRNA-disease association predictive accuracy is improved by coupling known and predicted miRNA-protein associations and protein-disease associations. Xu et al. [21] proposed the computational model based on the miRNA target dysregulated network (MTDN) to prioritize novel miRNAs related to specific disease. MTDN is constructed by combining miRNA-target association prediction and miRNA expression profile. In general, above computational models rely heavily on miRNA-target interactions. However, only less than 40% molecular bases of human diseases are partly known, which makes miRNA-target interactions inaccurate and limits the performance of those methods.

Xuan et al. [23] proposed the HDMP prediction method based on calculating the functional similarities of weighted k-neighbor miRNAs. Chen et al. [1] proposed Random Walk with Restart for MiRNA-Disease Association (RWRMDA), which first performs global analysis using similarity networks and miRNA-disease associations. HDMP and RWRMDA obtained high predictive accuracy according to cross validation results and case studies. However, they can't be accommodated to diseases without known associated miRNAs. Chen and Yan [3] further developed a semi-supervised method called Regularized Least Squares for MiRNA-Disease Association (RLSMDA). By incorporating disease similarity network, miRNA similarity network and miRNA-disease associations, RLSMDA can predict potential miRNAs related to diseases without any known related miRNA and negative samples are not needed. Chen et al. [2] proposed a novel WBSMDA method by integrating miRNA-disease associations and Gaussian interaction profile kernel similarities. Both RLSMDA and WBSMDA can be applied to diseases without any known related miRNA, but the predictive accuracy is still not very satisfactory.

Taken together, the limitations of previous methods are summarized as follows. Initially, some methods need negative samples that are uneasy to validate; furthermore, some methods can't predict potential miRNAs related to diseases without any known miRNA-disease relation; additionally, the incomplete datasets make some methods less accurate; finally, most computational models are closely coupled with specific type of data, and the performance suffers when datasets are expanded or updated. To solve above problems, we propose the Distributed and Privatized Framework for MiRNA-Disease Association prediction (DPFMDA), which is a distributed and extensible framework for miRNA-disease association prediction.

Considering the rapid growth of biomedical data, it is critical to develop an extensible and effective framework for improving accuracy of predicting potential miRNA-disease association. Firstly, DPFMDA independently extracts the miRNA and disease feature matrix factors from distributed datasets. Secondly, collaborative analysis is performed on clusters of matrix factors to generate reference factors. Thirdly, the reference factors are used as regularization terms for more accurate feature matrix estimation. The prediction accuracy is increased compared to state-of-the-art methods. Extensive experiments have been performed on DPFMDA. AUCs of 0.8859, 0.8267 and 0.8829 (+/-0.0013) are achieved for global LOOCV, local LOOCV and 5-fold cross validation respectively. DPFMDA is further evaluated with case studies on Esophageal Neoplasms, Prostate Neoplasms, Breast Neoplasms and Hepatocellular Carcinoma. As a result, 45, 44, 49 and 43 out of the top 50 predicted miRNA-disease associations for these diseases were confirmed by recent experimental reports.

2. Methods

We propose the Distributed and Privatized Framework for MiRNA-Disease Association prediction(DPFMDA). This computational model improves the prediction accuracy of the association between miRNA and disease. The subsections below explain the proposed method in more detail.

2.1. Data

First, in order to obtain the associations between miRNAs and diseases, we adopt three kinds of matrices as the input. These matrices denote human miRNA-disease associations, miRNA functional similarity and disease semantic similarity, respectively.

The HMDD database is used as the source of human miRNA-disease associations. There are 5430 experimentally verified records which partly reveal the connection between 383 kinds of diseases and 495 kinds of miRNAs in the latest version. Adjacency matrix A is constructed to describe the miRNA-disease associations. $A(d(i), m(j)) = 1$ means miRNA $m(j)$ is experimentally verified to be related to disease $d(i)$, otherwise the association is unconfirmed. We use variables n_m, n_d to denote the number of miRNAs and diseases in the database. Therefore, the goal of model is to predict the unconfirmed miRNA-disease associations with a high accuracy.

MiRNA functional similarity reflects the similarity of two miRNAs. Similar structures of miRNAs may have the similar functions that influence some kinds of diseases. The miRNA functional similarity scores are downloaded from http://www.cuilab.cn/files/images/cuilab/misim.zip. MiRNA functional similarity matrix is defined as FS. FS is a symmetric matrix, which has all 1s on its diagonal. The reasonable ranges of $FS(m(i), m(j))$ is between 0 and 1 which denotes the degree of similarity between miRNA $m(i)$ and miRNA $m(j)$. It's difficult to explore the actual relationship between the diseases by experiments. We use disease semantic similarity to represent the relationship. The semantic similarity of diseases can be obtained by two models. This paper introduces the first computational model of disease semantic similarity. The relations between diseases are obtained from MeSH database (http://www.ncbi.nlm.nih.gov/), then diseases can be described as $DAGs$. For a disease D, we define $DAG(D) = (D, T(D), E(D))$, where node set $T(D)$ contains the disease node itself and its ancestor nodes, and $E(D)$ is the direct edges from parent nodes to child nodes. The semantic value of disease D is calculated as follows:

$$DV1(D) = \sum_{d \in T(D)} D_D(d) \tag{1}$$

$$\begin{cases} D_D(d) = 1, & if\ d = D \\ D_D(d) = \max\{\Delta * D_D(d') | d' \in children\ of\ d\}, & if\ d \neq D \end{cases} \tag{2}$$

Where Δ is the semantic contribution fading factor. The more the diseases are different, the less semantic similarity between them. The semantic contribution diminishes as the increasing of distance between diseases in DAG. The contribution of disease to its own semantic value is 1.

Disease semantic similarity score measures the size of shared part between different disease $DAGs$. Semantic similarity matrix $SS1$ is defined as follows:

$$SS1(d(i), d(j))) = \frac{\sum_{t \in T(d(i)) \cap T(d(j))} (D_{d(j)}(t) + D_{d(j)}(t))}{DV1(d(i)) + DV1(d(j))} \tag{3}$$

$SS1$ is a symmetric matrix with all 1s on its diagonal. The values of $SS1(d(i), d(j))$ were scaled to the range [0, 1] which denotes the functional similarity score between disease $d(i)$ and disease $d(j)$. The second model is calculated in the same way as the literature [23]. Considering the model 1 for disease semantic similarity,

Please cite this article as: L. Chen et al., DPFMDA: Distributed and privatized framework for miRNA-Disease association prediction, Pattern Recognition Letters (2017), http://dx.doi.org/10.1016/j.patrec.2017.07.008