Auto-Encoder and Representation Learning Based MiRNA-Disease Association Prediction

Yutao Zhang[1], Xiya Lu[2*]

1. Fuzhou University, Fujian 350108, China.

2. Beijing Information Science and Technology University, Beijing 100101, China.

Abstract: As expressions of miRNAs are often associated with diseases, understanding the pathophysiology of illness at the miRNA level is beneficial for the treatment and prevention of associated diseases, as well as the creation of related medicines. Recent computational methods for predicting miRNA-disease associations integrate their pertinent heterogeneous data. The difficulty in this study is how to extract the implied associations from sparse data. In the present study, by drawing on natural language processing, a learning-based method is used to extract dense and high-dimensional representations of illnesses and miRNAs from integrated disease semantic similarity, miRNA functional similarity, and heterogeneous related interaction data. To predict disease-miRNA associations, we use a deep autoencoder and its reconstruction error as a measurement. Our experimental results suggest that our strategy is comparable to cutting-edge methods for predicting disease-related miRNAs.

Keywords: MiRNAs Auto-Encoder Gaussian Interaction Profile Kernel Function

1. Introduction

MicroRNAs are single-stranded non-coding tiny RNAs that are produced by the human body[1]. Recently miRNA has been discovered to be engaged in a number of biological processes, including cell proliferation, development, apoptosis, differentiation, and so on, thanks to the fast growth of genome sequencing technology and computer technology, according to quantities of studies. Understanding the pathophysiology of illness at the miRNA level is beneficial for the treatment and prevention of associated diseases, as well as the creation of related medicines.

Therefore, it's a popular topic about predicting miRNA and its potential disease subjects in the field of biomedicine. This has aided in the study of disease etiology, the development of novel medicines, and the formulation of customized diagnoses and treatments for a variety of complicated human diseases. Thus, we should build an accurate and efficient model in order to study the potential link between microRNA and disease. Computational prediction models not only rely on expensive and time-consuming biological experiments, but can also play a role by predicting the association of potential miRNAs with diseases and prioritizing candidate miRNAs. In addition, doing research about this topic can also create and promote befitting prediction models.

2. Research methods

2.1 Mirna-disease relationships

The Human MicroRNA Disease Database provided the miRNA-disease data used in this research (HMDD). The database contains thousands of experimentally validated miRNA illness correlations. With 5430 documented correlations, HMDD v2.0 contains 492 miRNAs and 329 illnesses. And there are 1206 miRNAs and 893 diseases in HMDD v3.2, with 35548 verified associations. We linked this research's data collection to a database of 492 miRNAs and 329 diseases for this investigation. We obtained 12030 known relationships between miRNAs and illnesses from HMDD after processing the data. We use a matrix A to represent disease and RNA. If there is a link between miRNA mi and disease dj, then Aij = 1, if there is no association or there may be an association but currently it is not found, then Aij=0. Therefore, the miRNA-disease association matrix is a two-dimensional matrix composed of 0 and 1.

2.2 Disease semantic similarity

The National Library of Medicine provided the mesh database, which includes numerous illness descriptions. The illness semantic similarity was calculated using a directed acyclic graph (DAG). We define A(D) and B(D) for the node D. The node set is A(D), while the edge set is B(D). Node D and its ancestor nodes are included in A(D), and B(D) reflects the direct relationship between the parent and child nodes. The disease semantic similarity matrix is defined by SS, and the similarity between diseases di and dj is denoted by SS(di, dj).

2.3 Functional similarity of mirna

Based on the notion that miRNAs with similar activity are frequently associated with similar illnesses. The similarity of two miRNAs can be assessed by analyzing the similarities of two diseases associated with miRNAs. In our research, we got the data directly and then utilized it to generate a 383x383 matrix MS. MS (mi, mj) represents each element in MS and displays the functional similarity of miRNA mi and mj. SD is the same size as SS and KD, with the exception that SD, a is the weight, which can range from 0 to 1. Variable values produce significantly diverse prediction outcomes, which will be explained.

2.4 Gaussian interaction profile kernel similarity

In our study, in order to transform the originally sparse and discontinuous data into dense and continuous data, we chose to use Gaussian interaction profile kernel function. Using known human miRNA–disease data, Gaussian similarity for both miRNAs and illnesses can be estimated. We constructed a matrix which is 383x383 and 492x492 respectively representing for MiRNA similarity and disease similarity. The Gaussian interaction kernel similarity of illnesses is represented by GD(di, dj). Similarly, the miRNA Gaussian profile interaction kernel similarity is denoted by GM(mi, mj).

2.5 Representation

In this article, in order to further make the data dense for extracting the implied associations, we drew on techniques from natural language processing to train two models to represent diseases and miRNA by learning vectors of representation, using the word embedding method.

Firstly we numbered each disease and miRNA with a unique number. it was utilized as a search to obtain vector di from an embedding matrix d, with each row indicating a different miRNA or illness. d is initialed at random. and then, after a number of epochs of training, we can obtain the characteristics of the specified dimension.

The illness semantic similarity is a matrix that is sparse, and utilizing it alone is challenging to obtain appropriate prediction performance. furthermore, the kernel similarity of the gaussian interaction profile GD is estimated using known human miRNA–disease relationships, which is insufficiently precise. so, to get great prediction performance, it is essential to combine the disease semantic similarity ds with the gaussian interaction profile kernel similarity gd. we just let them add up linearly like:

$$DSS(d_i,d_j) = aDS(d_i,d_j) + (1 - a)GD(d_i,d_j)$$

DSS has the same dimensions as DS and GD. DSS has the same dimensions as DS and GD. And the weight, whose value is between 0 and 1. Varied values result in dramatically various prediction results. According to earlier research, each component in matrices DS and GD is in the range [0,1]. Our approach takes $DSS(d_i, d_j)$ as the distance between two illnesses, di and dj. Cosine similarity can be used to compare the similarity of two vectors. that is extensively used in data mining and information retrieval. It is based on the cosine of the angle between them. We employed cosine similarity to assess the distance between two illnesses, and DSS is used as labels for ground-truth. In order to get a non-negative within the range of [0,1], we calculate it in the following way:

$$DSSD(d_i,d_j) = \frac{1}{2} + \frac{1}{2}\frac{d_i \cdot d_j}{||d_i||d_j||}$$

Here, di is the vector that represents the illness, di. SD0 represents the calculated distance between illnesses di and dj, while SDdi represents the ground-truth similarity. To learn disease representation D, we built a regression model. In high-dimensional disease spaces, two vectors with a high similarity score are significantly comparable. As indicated below, the di disease model seeks to reduce loss across all samples:

$$argmin\frac{1}{N_d}\sum^{N_d}||DSSD(d_i,d_j)^2 - DSS(d_i,d_j)^2||$$

where the number of training samples is denoted by Nd. At each training iteration, as a criterion, the mean squared loss is used, and the disease matrix D is updated using the stochastic gradient descent (SGD) approach with backpropagation.As for the representation of miRNA, we use the same method.

2.6 Integration strategies and predicting

To begin, we use the Word Embedding function to vectorize the words. The disease representation vector and miRNA representation vector processed by the embedding function are spliced into disease-miRNA pairs. Then we utilize an autoencoder, which is an unsupervised learning approach initially developed by Hinton in the 1980s.There are two components to our autoencoder: encoding and decoding. The input is compressed into a low-dimensional vector in the encoding phase, and the decoding part restores the vector to the original input as the output. The backpropagation algorithm makes the two parts the same.

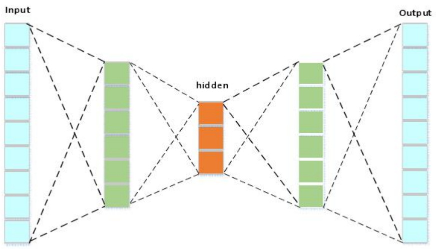

Figure 1 Auto-encoder process diagram

Finally, our autoencoder's loss is the total of all restructuring mistakes from all training samples, and it is stated as follows:

$$l(x,\tilde{x}) = \sum_{i=1}^{N} j(x_i - \tilde{x}_i)j^2 + \lambda j J_{\mathrm{h}}(x_i)j^2$$

The number N represents the number of known miRNA and disease relationships. The squared loss is the first loss item, the hyperparameter is λ. The deep autoencoder is trained to reduce the aforementioned loss, and autoencoder parameters are updated on a regular basis.

3. Results

3.1 The data information

The Human microRNA Disease Database provided us with the mi-RNA-disease data (HMDD). There are 1206 miRNAs and 893 diseases in HMDD v3.2, with 35548 verified associations. We linked the data collected in this investigation to a database containing 492 miRNAs and 329 diseases. And we found 12030 known relationships between miRNAs and illnesses from HMDD after analyzing the data. Then we gained the mesh database from the National Library of Medicine. The illness semantic similarity was calculated using a directed acyclic graph(DAG).

3.2 Analysis and comparison

To correctly measure the model's prediction performance, use the average AUC, AUPR, precision, recall, and F1 of 10 5-fold cross-validation as assessment markers, where The area under the ROC curve with FPR as the abscissa and TPR as the ordinate is known as AUC, whereas the area under the PR curve is denoted as AUPR, with recall as the abscissa and accuracy as the ordinate:

$$TPR = TP \,/\, (TP + FN)$$

$$FPR = FP \,/\, (FP + TN)\; recall = TP \,/\, (FP + TN)$$

$$precision = TP \,/\, (TP + FP)$$

Figure 1 Comparison of AUC values of different methods in three diseases

Types	WBSMDA	HDMP	MIDP	Our method
breast cancer	0.7538	0.7959	0.8057	0.8262
lung cancer	0.8002	0.9059	0.8924	0.9266
colon cancer	0.7053	0.8120	0.7971	0.8704

Table 1 shows the experimental outcomes. It indicates that our method has the highest AUC in lung cancer, reaching 0.9266; The average AUC values of MIDP, WMSMDA and HDMP algorithms are 0.8317, 0.7531, 0.8081 and 0.8744.The performance is slightly better than other algorithms.

WBSMDA, miRNAs were discovered by combining known miRNA-disease connections, miRNA functional similarity, disease semantic similarity, and Gaussian interaction profile kernel similarity. The final score for putative miRNA-disease association inference was computed by combining the Within-Score and Between-Score[2]. HDMP, people created a computer model of human disease-related miRNA prediction (HDMP) by taking into account each miRNA's k most comparable neighbors. To generate more trustworthy relevance ratings for the unlabeled miRNAs, the k closest neighbors of each miRNA and miRNA functional similarity were merged. Furthermore, HDMP gave more weight to miRNAs that were part of the same miRNA family or cluster[3]. MIDP, people created a computer model of human disease-related miRNA prediction (HDMP) by evaluating each miRNA's k most comparable neighbors. The k closest neighbors of each miRNA and miRNA functional similarity were used to generate more trustworthy relevance ratings for the unlabeled miRNAs. Furthermore, HDMP gave more weight to miRNAs in the same miRNA family or cluster[4].

In the miRNA-disease association prediction task, our method was compared with DNN,RWR_DNN and MCMDA. The analysis of AUROC, AUP-R, Precision and F1-score shows that our method is slightly better than others.

DNN, DNNs map input data into a low-dimensional space using many layers of non-linear functions, encapsulating highly non-linear network structure in efficient low-dimensional characteristics. DNN's multi-layer design is essential for learning richer network representations[5]. RWR, for each miRNA-disease combination, people mapped the disease's causative genes and miRNA target genes into the PPI network. The random walk with restart (RWR) technique was then used to generate a gene rank list. In the above-mentioned gene list, each miRNA target gene was assigned a likelihood value. The higher the likelihood value, the closer the miRNA target gene was to a known illness gene[6]. MCMDA, based on the known miRNA-disease connections, people presented a matrix completion technique for MDA (MCMDA). For predicting possible connections, MCMDA used the matrix completion technique to update the adjacency matrix of known miRNA-disease relationships[7].

4. Discussion

Prediction of miRNA-disease association is a research hotspot in recent years, which has far-reaching significance for revealing the mechanism of complex diseases at the molecular level, so it has important research value. Most of the methods used to calculate and forecast disease-related miRNA make use of similarity data, and some approaches do not make full use of biological data in similarity calculation, thus the use of known correlation data is not explored thoroughly. However, major challenges in current research include small sample sizes, a lack of negative sample data, and related miRNA prediction of new illnesses.

In this research, we used the Gaussian interaction profile kernel function to transform sparse and discontinuous data into dense and continuous data. And then we made use of natural language processing techniques to train two models to describe illnesses and miRNA by learning vectors in order to make the data more continuous. We integrated by using Word Embedding and Autoencoder. In fact, our results is similar to other methods.

According to existing research, while this work analyzes and explores the prediction of disease-associated miRNA and addresses certain difficulties, there are still some issues to be resolved and opportunity for development at this point.

Firstly, The method in this paper does not use multiple sources of miRNA or disease-related data, but only based on the miRNA-disease association, disease semantic similarity, functional similarity of miRNA, and Gaussian interaction kernel similarity are estimated. Maybe, some gene expression data as well as topological information.

What's more, integrating two similarities using a regression model is not necessarily a good approach in my opinion. It's better to use some other methods such as autoencoder for multi-source data integration[8].

Due to time limitations, we will try these possibilities in our future studies and research.

5. Tables and figures

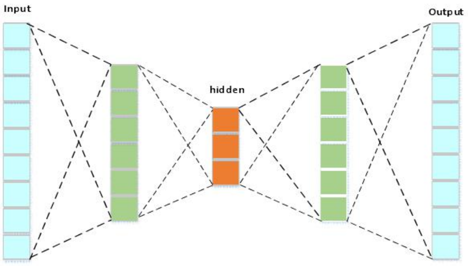

Figure 1 Auto-encoder process diagram

Table 1 Comparison of AUC values of different methods in three diseases

Types	WBSMDA	HDMP	MIDP	Our method
breast cancer	0.7538	0.7959	0.8057	0.8262
lung cancer	0.8002	0.9059	0.8924	0.9266
colon cancer	0.7053	0.8120	0.7971	0.8704

Table 2 5-fold cross validation results of different models

Types	AUROC	AURP	Precision	Recall	F1-score
DNN	0.9322	0.9279	0.8688	0.8101	0.8324
RWR_DNN	0.9199	0.9203	0.8380	0.8507	0.8442
MCMDA	0.9229	0.9315	0.8483	0.8697	0.8587
Our method	0.9420	0.9408	0.8720	0.8987	0.8736

References

[1] Wang, X Z., 2020. A Research on Prediction Method of Disease Related-miRNA Based on Biased Heat Conduction Recommended Algorithm.

[2] Chen, X., Chenggang ClarenceYan, et al, 2016. WBSMDA: Within and Between Score for MiRNA-Disease Association prediction.

[3] P. Xuan, K. Han, M. Guo, Y. Guo, J. Li, J. Ding, Y. Liu, Q. Dai, J. Li, Z. Teng, 2013 Prediction of microRNAs associated with human diseases based on weighted k most similar neighbors, PLoS One 8.

[4] P. Xuan, K. Han, Y. Guo, J. Li, X. Li, Y. Zhong, Z. Zhang, J. Ding. 2015 Prediction disease-associated microRNAs based on random walk.

[5] EdgarManzanarez-Ozuna, Dora-Luz Flores, EverardoGutiérrez-López, David Cervantes and Patricia Juárez,2018. Model based on GA and DNN for prediction of mRNA-Smad7 expression regulated by miRNAs in breast cancer.

[6] Liao, F., Chen, X., Peng, PL., Dong, WG., 2020, RWR-algorithm-based dissection of microRNA-506-3p and microRNA-140-5p as radiosensitive biomarkers in colorectal cancer.

[7] Li, J.-Q. et al. 2017 MCMDA: matrix completion for MiRNA–disease association prediction. Oncotarget, 8, 21187–21199.

[8] Vladimir Gligorijević, Meet Barot, Richard Bonneau, deepNF: deep network fusion for protein function prediction, Bioinformatics, Volume 34, Issue 22, 15 November 2018, Pages 3873–3881.