Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Subject Section

A learning-based framework for miRNA-disease association prediction using neural networks

Jiajie Peng [1,2,3], Weiwei Hui [1], Qianqian Li [1], Bolin Chen [1,2,3], Zhongyu Wei [4,*] and Xuequn Shang [1,2*]

[1]School of Computer Science, Northwestern Polytechnical University, Xi'an, 710072, China and
[2]Key Laboratory of Big Data Storage and Management, Northwestern Polytechnical University, Ministry of Industry and Information Technology, Xi'an, 710072, China and
[3]Centre for Multidisciplinary Convergence Computing (CMCC), School of Computer Science, Northwestern Polytechnical University, Xi'an, 710072, China and
[4]School of Data Science, Fudan University, Shanghai, 200433, China.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: A microRNA (miRNA) is a type of non-coding RNA, which plays important roles in many biological processes. Lots of studies have shown that miRNAs were implicated in human diseases, indicating that miRNAs might be potential biomarkers for various types of diseases. Therefore, it is important to reveal the relationships between miRNAs and diseases/phenotypes.
Results: We proposed a novel learning-based framework, MDA-CNN, for miRNA-disease association identification. The model first captures richer interaction features between diseases and miRNAs based on a three-layer network with an additional gene layer. An auto-encoder is then employed to identify the essential feature combination for each pair of miRNA and disease automatically. A final label is given by a convolutional neural network taking the reduced feature representation as input. The evaluation results showed that the proposed framework outperformed some state-of-the-art approaches in a large margin on both tasks of miRNA-disease associations prediction and miRNA-phenotype associations prediction.
Availability: The software will be available after the manuscript is accepted.
Contact: jiajiepeng@nwpu.edu.cn;zywei@fudan.edu.cn;shang@nwpu.edu.cn
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

A microRNA (miRNA) is a type of non-coding RNA containing about 22 nucleotides and it has been found in different types of organisms, such as plants, animals, viruses (Liu *et al.*, 2015; Meister and Tuschl, 2004; Huang *et al.*, 2011; Li *et al.*, 2013a; Yuan *et al.*, 2015), etc. Existing researches revealed that miRNAs played important roles in different biological processes, including cell growth (Ambros, 2003), cell cycle regulation (Carleton *et al.*, 2007), immune reaction (Taganov *et al.*, 2006), tumor invasion (Ma *et al.*, 2007) and cell fate transitions (Shenoy and Blelloch, 2014) by targeting specific messenger RNAs (mRNAs) and regulating gene expression and mRNA degradation (Shi *et al.*, 2013; Jiang *et al.*, 2010).

Studies also indicated that many miRNAs were implicated in human diseases (Jiang *et al.*, 2008; Li *et al.*, 2013b), e.g., cancer (Volinia *et al.*, 2012), viral diseases (Huang *et al.*, 2007), immune-related diseases (Huang *et al.*, 2007), Parkinson's disease (Kim *et al.*, 2007), etc. According to the statistics of HMDD (the Human microRNA Disease Database) (Li *et al.*, 2013b), there are more than 10,000 identified miRNA-disease associations, involving 572 miRNAs and 378 diseases. Moreover, recent experiment results (Goulart *et al.*, 2015; Dweep and Gretz, 2015) showed strong links between miRNAs and phenotypes. Considering miRNAs might be potential biomarkers for various types of diseases, it is important to further explore the relationships between miRNAs and diseases/phenotypes to understand pathogenicity mechanisms.

© The Author 2015. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

Biological experimental verification of associations between a pair of miRNA and disease/phenotype is time-consuming and costly in terms of financial input. Therefore, researchers explored to predict miRNA-disease association automatically based on computational approaches (Pallez *et al.*, 2017; Zou *et al.*, 2015; You *et al.*, 2017). Most of existing approaches are based on network. The basic assumption is that miRNAs associated with the same or similar diseases are more likely to be functionally related (Chen *et al.*, 2012), and the crucial technical point is similarity computation (Zou *et al.*, 2015) for miRNA-miRNA, disease-disease and miRNA-disease associations (Zou *et al.*, 2015). According to what information is involved in similarity computation, network-based approaches can be loosely grouped into two categories (Zou *et al.*, 2015), local network similarity methods (Kertesz *et al.*, 2007; Lewis *et al.*, 2003; Wei *et al.*, 2012; Xuan *et al.*, 2013; Han *et al.*, 2014) and global network similarity methods (Chen *et al.*, 2012; Shi *et al.*, 2013).

The simple idea of measures in the local network similarity-based category is that similar miRNAs might be associated with similar diseases. For example, Jiang et al. (Jiang *et al.*, 2010) proposed a hypergeometric distribution-based method to identify the miRNA-disease association based on miRNA-miRNA and disease-disease network. HDMP method was proposed to predict miRNA-disease associations based on the most similar neighbors (Xuan *et al.*, 2013). Specifically, the association score between a miRNA mr and a disease d was calculated by considering the association between d and the k most similar neighbors of mr. The evaluation results showed that HDMP performed better than Jiang's method. Local network similarity-based methods only consider the direct edge information contained in the involved networks, neglecting the global structure of these networks.

In the global network similarity-based category, a few proposed methods adopted the random walk with restart (RWR) model (Chen *et al.*, 2012; Shi *et al.*, 2013) to predict miRNA-disease associations. For example, RWRMDA adopted RWR on a miRNA-miRNA association network to measure the similarities between known disease-associated miRNAs and candidate miRNAs (Chen *et al.*, 2012). In addition, Shi et al. applied RWR on a protein-protein interaction network to measure the similarities between genes linked to a disease and target genes of candidate miRNAs (Shi *et al.*, 2013). A key step of the global network similarity-based methods is mapping miRNAs and diseases to the same network. Therefore, the result may highly depend on the quality of networks constructed. Chen et al. proposed a method named RLSMDA to predict the miRNA-disease association using a semi-supervised method that does not need negative samples (Chen and Yan, 2014).

Recently, PBMDA constructed a heterogeneous graph including miRNA-miRNA, miRNA-disease and disease-disease relations to calculate the association score between miRNAs and diseases (You *et al.*, 2017). Experiment results showed that PBMDA performed better than other four recently proposed methods (HDMP (Xuan *et al.*, 2013), RLSMDA (Chen and Yan, 2014), WBSMDA (Chen *et al.*, 2016) and RWRMDA (Chen *et al.*, 2012)).

Although many attempts have been made to identify miRNA-disease association automatically, most of them are un-supervised ones solely based on networks without using any labeled information. This largely limits the performance of state-of-the-art systems. Instead of using networks to compute association scores directly, we explore to use them extract interaction features for miRNA-disease pairs and predict their associations in a supervised fashion.

In general, miRNA-disease networks contain two levels, one for miRNAs and the other for diseases. Similarities are computed within the same level and cross levels. We argue that similarity computation using target concepts (miRNA and disease) directly is unable to catch the deep interaction patterns between them. As we know, miRNAs are implicated in many human diseases by regulating their target genes (Ardekani and Naeini, 2010), indicating that gene plays a crucial role to bridge miRNAs and diseases. Therefore, we believe gene information should be considered in studying miRNA-disease associations. To address this problem, we introduce a gene layer in the middle of the miRNA and disease layers to form a three-layer-network and propose a novel way to extract interaction features for miRNA-disease pairs.

In past a few years, deep neural networks (e.g. convolution neural networks) have been applied to many bioinformatics applications (Min *et al.*, 2017) and have produced promising results. It has been proved that convolutional neural networks obtain the ability to detect meaningful combinations of features from the input data automatically (Angermueller *et al.*, 2016). Considering features extracted from three-layer network might contain noises, we propose to use convolutional neural network to learn the best combination of features and predict the final labels for a given miRNA-disease pair.

In this article, we present a novel learning-based framework, MDA-CNN, to identify the association between a pair of miRNA and disease. Here are four major contributions:

- We introduce a learning-based framework for the task of miRNA-disease association prediction that contains three components, namely network-based feature extractor, auto-encoder-based feature selector and CNN-based association predictor.
- To better represent the correlation between miRNAs and diseases, we construct a three-layer network with an additional gene layer in the middle. A novel feature representation is then proposed based on the regression model.
- We employ a deep convolutional neural network (CNN) architecture to deal with the feature vectors produced in the previous step to determine the final label for miRNA-disease pairs.
- The evaluation results show that MDA-CNN outperforms all the other state-of-the-art systems for both miRNA-disease and miRNA-phenotype association identification.

2 Methods

We propose a novel algorithm called MDA-CNN to predict miRNA-disease association. To achieve the goal, MDA-CNN contains three steps. The framework of MDA-CNN is shown in figure 1. First, given a three-layer network (see Figure 1a), we applied a regression model to calculate the disease-gene and miRNA-gene association scores. The disease-gene association scores are calculated based on disease layer and gene layer. The miRNA-gene association scores are calculated based on miRNA layer and gene layer. For each miRNA or disease, we generated a vector to represent its feature based on its associations with genes (Figure 1b). Second, given a pair of miRNA and disease, a concatenated vector of miRNA and disease was passed through a auto-encoder-based model to obtain a low dimensional representation of a miRNA-disease pair (Figure 1c). Third, a deep convolutional neural network (CNN) architecture was constructed to predict the association between miRNA and disease based on the vector space representation obtained in last step (Figure 1d).

2.1 Network-based feature extraction

As we know, miRNAs regulate gene expression post-transcriptionally. In this work, we used the gene-layer network as the bridge for feature representation, which can be easily used for miRNA-disease association prediction. To achieve this goal, given a disease d or miRNA m, we need to calculate the association score between d (or m) and every gene in the gene-layer network. In the following, we take disease and gene layer as the example to illustrate our algorithm. By integrating a network N_d based on disease similarities, a protein interaction network N_g

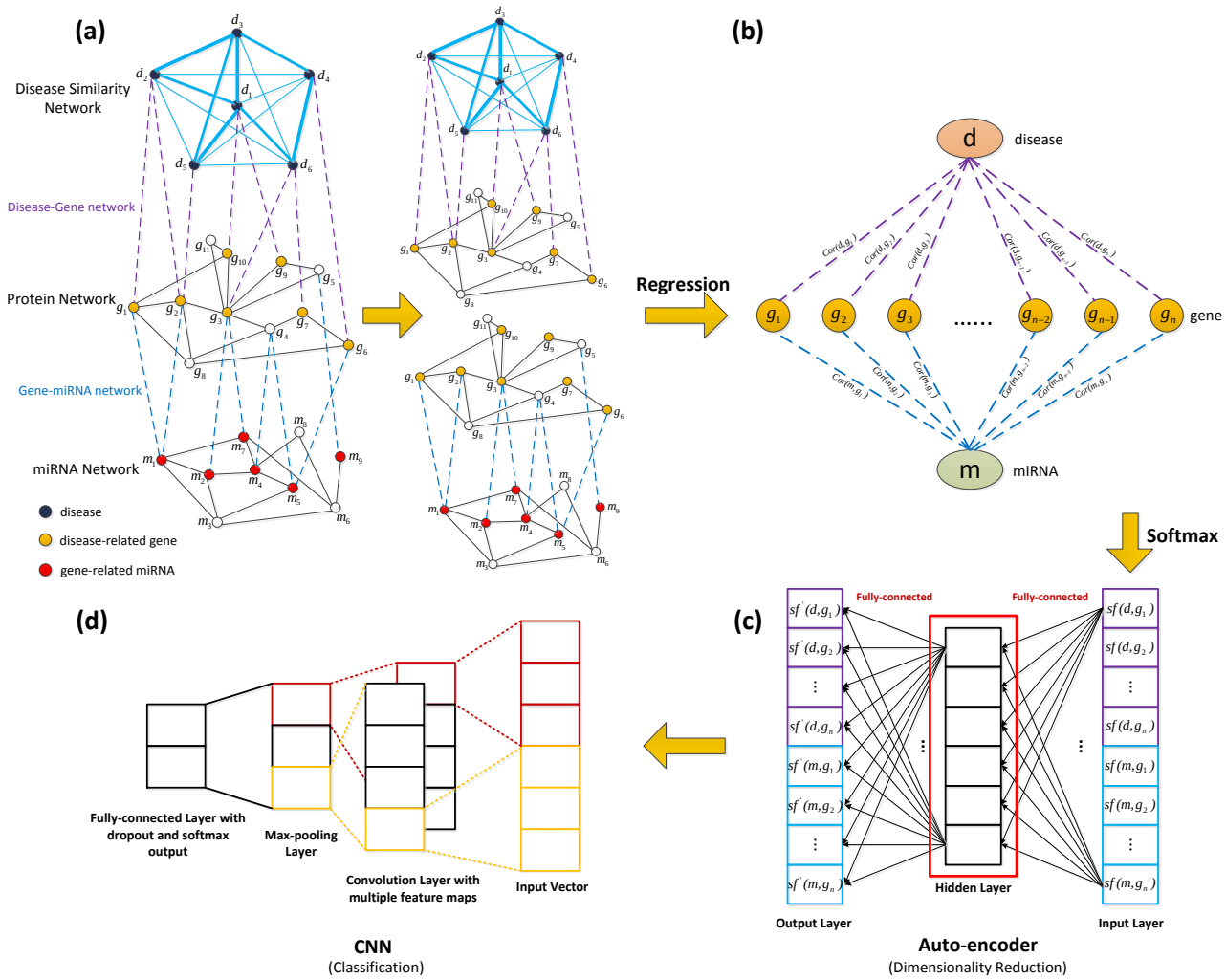

Fig. 1. The workflow of MDA-CNN.

and known disease-gene associations A_{dg}, we adopt a two-layer network regression model (Wu *et al.*, 2008) for association score calculation. The basic assumption is that phenotypically similar diseases may be associated with functionally related genes. Based on N_d, N_g and A_{dg}, the association score between gene g and disease d can be defined as follows:

$$R(g,d) = \sum_{g^{'} \in G(d)} e^{-dis(g,g^{'})} \tag{1}$$

where $G(d)$ is the set of genes associated with disease d in the protein interaction network N_g. $dis(g,g^{'})$ is the topological distance between g and $g^{'}$ in the protein interaction network N_g. Specifically, $dis(g,g^{'})$ is calculated as $sdis(g,g')^2$, where $sdis(g,g')$ is the shortest path between g and $g^{'}$ in the protein network N_g. Furthermore, given two diseases d_1 and d_2, the similarity between d_1 and d_2 is defined as follows:

$$Sim(d_1,d_2) = C_{d_1} + \sum_{g_1 \in G(d_1)} \beta_{d_1 g_1} R(g_1,d_2) \tag{2}$$

where $\beta_{d_1 g_1}$ is the regression coefficient of this linear regression model, $G(d_1)$ is the set of genes associated with d_1 in protein interaction network N_g, C_{d_1} is a constant. The similarity score between d_1 and d_2 is calculated based on the relationship between gene sets associated with d_1 and d_2 respectively.

Given a disease d_1, the original association score between d_1 and other diseases in N_d is obtained from the edge weight in N_d. For every gene g related with d_1, $R(g,d_2)$ is calculated based on Equation 1. Therefore, the regression model (Equation 2) can be trained for each disease d_1. After the regression model is trained for each disease, we could update the similarity between two diseases based on Equation 2. Given a disease d_1 and a gene g, their correlation score can be calculated as follows:

$$Cor(d_1,g) = \frac{cov(S_{d_1},R_g)}{\delta(S_{d_1})\delta(R_g)} \tag{3}$$

where $cov()$ and $\delta()$ represents covariance and standard deviation respectively, S_{d_1} is a vector containing association scores between d_1 and each disease in N_{d_1} (see Equation 4).

$$S_{d_1} = [Sim(d_1,d_1), Sim(d_1,d_2), ..., Sim(d_1,d_n)] \tag{4}$$

R_g is a vector containing association scores between g and each disease in N_{d_1} (see Equation 5).

$$R_g = [R(g,d_1), R(g,d_2), ..., R(g,d_n)] \tag{5}$$

Based on Equation 3, given a disease d_1, we can calculate its correlation scores with all genes involved in N_g.

Let N_m, N_g and A_{mg} be a miRNA network based on miRNA similarities, a protein interaction network N_g and known miRNA-gene associations crossing the two networks. Using the aforementioned model, given a miRNA m_1, we can calculate correlation scores with all genes involved in N_g.

Feature representation is the key step before applying the machine learning algorithm. However, to the best of our knowledge, there is no feature representation method in the miRNA-disease relation prediction problem. By adding the gene layer, we can generate a vector to represent features of miRNA and disease.

Given a disease d, we calculated correlation scores between d and each gene involved in the gene layer based on Equation 3. After that, a vector can be generated as

$$X_d = [Cor(d, g_1), Cor(d, g_2), ..., Cor(d, g_n)] \tag{6}$$

where g_i represents a gene involved in N_g, n is the number of genes involved in N_g. To reduce the influence of extreme values (outliers) in X_d, we applied softmax normalization (Grover and Leskovec, 2016) on X_d. Specifically, the normalized vector X'_d is represented as follows:

$$X'_d = [\frac{exp(Cor(d, g_1))}{\sum_{i=1}^{n} exp(Cor(d, g_i))}, ..., \frac{exp(Cor(d, g_n))}{\sum_{i=1}^{n} exp(Cor(d, g_i))}] \tag{7}$$

Similarly, given a miRNA m, a vector X'_m is generated as follows:

$$X'_m = [\frac{exp(Cor(m, g_1))}{\sum_{i=1}^{n} exp(Cor(m, g_i))}, ..., \frac{exp(Cor(m, g_n))}{\sum_{i=1}^{n} exp(Cor(m, g_i))}] \tag{8}$$

For each disease or miRNA, we can generate a vector for feature representation based on gene layer information.

2.2 Auto-encoder-based feature selection

The length of vector generated in previous step (e.g. X'_d and X'_m) is the same as the number of genes included in network N_g, which might be very large and noisy. In extreme case, the length of the vector is the same as the number of known genes in human genome. Therefore, we applied auto-encoder to identify the essential feature combination and reduce the dimension for each pair of miRNA and disease automatically.

Auto-encoder is used to do dimensionality reduction work for the downstream machine learning task such as classification, visualization, communication and storage of high-dimensional data (Bourlard and Kamp, 1988; Hinton and Salakhutdinov, 2006; Chicco *et al.*, 2014; Ranzato *et al.*, 2007). Unlike the widely-used method, e.g. principal components analysis (PCA), auto-encoder is a non-linear generalization of PCA that uses an adaptive "encoder" network to transform the high-dimensional data into a low-dimensional code, and a similar "decoder" network to recover the data from the low-dimensional code. The low-dimensional code forces the network to compress the input data into a low dimensional representation. The detail of auto-encoder is described in the supplementary document.

In our experiments, the vector X'_d and X'_m are concatenated into a single vector before feeding into the auto-encoder model. Let n be the number of genes involved in the network N_g. The original dimension of input is $2n$. In our model, we use the mean-square error (Wax and Ziv, 1977) (MSE) in the loss function. The sigmoid activation function (Hosmer and Lemeshow, 2000) and Adam algorithm (Kingma and Ba, 2015) are used to optimize the MSE loss. Our auto-encoder network is trained by using the backpropagation algorithm (BP) algorithm (Rumelhart *et al.*, 1988). More detail about the parameters used in the model is described in the Results section.

2.3 Convolutional neural network-based association prediction

Convolutional neural network (CNN) was proposed in the late 1980s by Lecun (Lecun *et al.*, 1989) which performs very well in image classification (Krizhevsky *et al.*, 2012; Lecun *et al.*, 2010), face recognition (Lawrence *et al.*, 1997), sentence classification (Kim, 2014) and classification task on the structured-graph data (Atwood and Towsley, 2015). It is with this insight that we proposed a convolutional neural network for the task of miRNA-disease association prediction.The structure of the proposed model is shown in Figure 2. Our model includes

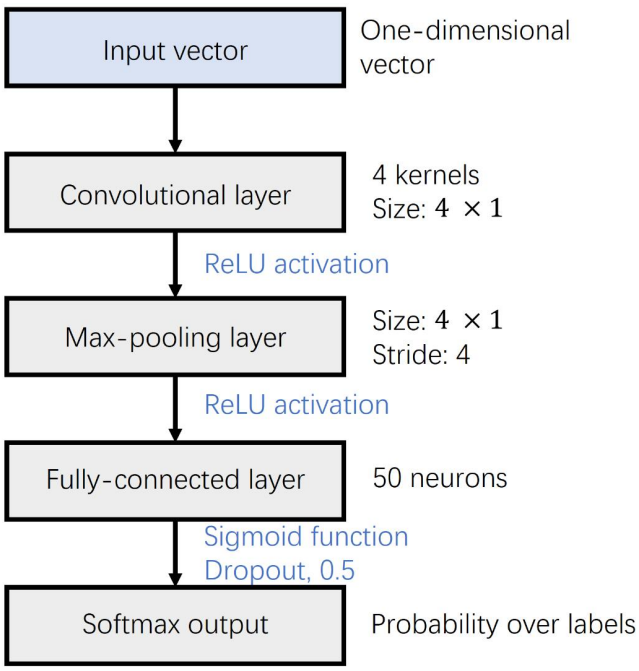

Fig. 2. The structure of proposed convolutional neural network model. The input is a one-dimensional vector. Our model consists of convolutional layer, max-pooling layer and FC layers. The output is the probability distribution over labels for each sample.

the following layers: convolution and activation layer, max-pooling layer, fully-connected layer and softmax layer. A convolutional layer and a rectified linear unit (ReLU, (Nair and Hinton, 2010)) activation layer are used to extract features from the one-dimensional input that was the output of dimension reduction step (see previous subsection). A pooling layer is used for dimensionality reduction. The final fully-connected layer and softmax layer are for the classification task.

The convolutional layer is responsible for learning subspace features of the input. The convolutional layer of our model consists of 4 kernels. 4×1 weight vector is convolved with the input vector with length L. After the convolution, for each kernel, we can obtain a feature map C (a particular feature extracted from the input) that is a one-dimensional vector with the length $(L - 4) + 1$. The feature map C is extracted by the following equation:

$$C_u = \sum_{j=1}^{4} W_j X_{u+j} \tag{9}$$

where $u \in \{0, ..., L - 4\}$, X is the input vector, and $W \in R^4$ is a one-dimensional weight vector, which is initialized by a truncated normal distribution with mean zero and standard deviation 0.1. A high C_u indicates that the kernel captures the features of the sub-region of the input very well. C_u is then passed through a ReLU function $f(x) = max(0, x)$ that

ignores negative outputs and propagates positive outputs from previous layer. Although there are various non-linearities, ReLU activations are the most popular because of its computational efficiency, sparsity and reduced likelihood of vanishing gradient (Nair and Hinton, 2010; Krizhevsky *et al.*, 2012; Lecun *et al.*, 2015).

Max-pooling layer is used to down-sample the latent representation after the convolutional layer. Max-pooling takes the maximum value over non-overlapping sub-regions (i.e., the pooling size is 4) of the output of convolutional layer and outputs the most important feature over a neighborhood in each feature map. Given an input sequence $f(C_u)$ ($u \in \{0, ..., L-4\}$), the output of pooling layer is shown as follows:

$$\max_{u\in\{0,\ldots,3\}} f(C_u), \max_{u\in\{4,\ldots,7\}} f(C_u), ..., \max_{u\in\{L-7,\ldots,L-4\}} f(C_u) \tag{10}$$

with length of $\frac{(L-4)+1}{4}$.

Convolutional layer and max-pooling layer can extract the important features from the input vector (Parameters: Number of kernels: 4, Filter size: 4×1, Pooling size: 4×1, strides: 4). The outputs of all the kernels are then concatenated into a one-dimensional vector and fed to the fully-connected layer.

The final two layers are a fully-connected layer and a softmax layer. There are 50 hidden units in the fully-connected layer. The output of pooling layer is $y \in R^n$, where n is the length of concatenated outputs of the pooling layer. The output of the fully-connected layer is :$f(W * y)$, where $W \in R^{n\times 50}$ is the weight matrix, and f is the ReLU activation. The final softmax layer is used for classification task. The predicted label probability of each sample is: $S(v^T f(W * y)) \in (0, 1)$, where $v \in R^{50}$ and $S(x)$ is the sigmoid function $S(x) = \frac{1}{1+e^{(-x)}}$.

3 Results

3.1 Data preparation

To evaluate the performance, MDA-CNN was tested on two tasks: miRNA-disease association prediction and miRNA-phenotype association prediction.

The disease similarity network was obtained from You's paper (You *et al.*, 2017). 383 diseases are involved in the disease similarity network. The known disease-gene associations were obtained from DisGeNET database (http://www.disgenet.org/web/DisGeNET/menu). We only used the manually curated disease-gene associations in our experiment. The protein network of human genes was downloaded from Human Protein Reference Database (HPRD) database (Baolin and Bo, 2007). We used a miRNA similarity matrix generated in (You *et al.*, 2017) to construct miRNA network. A threshold was set to remove the associations with low confidence. Finally, a miRNA network including 243 miRNAs and 1618 associations between miRNAs was generated by setting threshold as 0.6. The known miRNA-gene associations were downloaded from miRWalk2.0 database (Dweep and Gretz, 2015). Phenotype-gene associations were obtained from Human Phenotype Ontology (HPO) official website (http://human-phenotype-ontology.github.io/). The phenotype network was generated from the phenotype similarities calculated based on HPO. We use the method proposed in (Masino *et al.*, 2014) to calculate HPO-based similarity. Let p_1 and p_2 be two phenotypes, and S be the set of all common ancestors of p_1 and p_2. p_m is the phenotype has a minimum of gene annotations in S. The similarity between p_1 and p_2 is defined as follows.

$$sim(p_1, p_2) = -log\frac{N_{p_m}}{N} \tag{11}$$

Where N_{p_m} is the number of gene annotations of p_m (including gene annotations of its descendants) and N is the total number of genes involved in HPO. The validated miRNA-disease associations was downloaded from HMDD (Li *et al.*, 2013b). The validated miRNA-phenotype associations were also obtained from miRWalk2.0 database (Dweep and Gretz, 2015).

In the performance evaluation on miRNA-disease association prediction, we remove the diseases that have no related genes involved in the protein network. Similarly, we removed the genes that have no related miRNAs. 204 diseases, 1789 genes and 243 miRNAs are involved in our experiment. Finally, the number of validated miRNA-disease associations involved in the test is 2449. The 2449 validated miRNA-disease associations are used as the positive set. Because there was no database contained the unrelated miRNA-disease associations, we randomly generated 2449 (the same size of the positive set) pairs of miRNA-disease associations as the negative set. We used ten-fold cross validation (Kohavi *et al.*, 1995) to train and test our model.

In the performance evaluation of miRNA-phenotype association prediction, we removed the phenotypes that have no related genes involved in the protein network. Finally, 4238 phenotypes are involved in our experiment. Finally, the number of validated miRNA-phenotype associations involved in the test is 44280. These validated miRNA-phenotype associations are considered as the positive set. Because there was no database collected the unrelated miRNA-phenotype associations, we randomly generated 44280 (the same size of the positive set) pairs of miRNA-phenotype associations as the negative set. We also used ten-fold cross-validation to train and test our model.

3.2 Model parameters

The parameters of generated auto-encoder model are as follows. The original dimensions of miRNA-phenotype and miRNA-disease are 3578. The generated network has one hidden layer with 1024 units. For miRNA-phenotype and miRNA-disease association prediction, there are 256 and 128 samples for every batch respectively. The number of epochs are 10 and 90 for miRNA-phenotype and miRNA-disease association prediction respectively. The auto-encoder network is trained using the BP algorithm (Rumelhart *et al.*, 1988).

The parameters of generated convolutional neural network model are as follows. For the miRNA-disease association prediction task, we added a dropout layer (Hinton *et al.*, 2012) and employed l_2-norms of the weight vectors. In our model, the dropout percentage equals 0.5 and parameter of l_2- norms is 0.002. With 64 samples for every batch, we run 150 epochs. We used Adam algorithm with an initial learning rate equals $1e-2$ to optimize the loss on the whole network which is the sum of cross-entropy loss and l_2 loss. For miRNA-phenotype association prediction task, the batch size is 128 and the number of epochs is 10. We use Adam optimizer algorithm with an initial learning rate $1e-2$ to optimize the cross-entropy loss (Boer *et al.*, 2005).

3.3 Performance evaluation on predicting miRNA-disease associations

We evaluated the performance of MDA-CNN and the other three methods (i.e. CIPHER (Wu *et al.*, 2008),PBMDA (You *et al.*, 2017) and WPSMDA) on the task of predicting miRNA-disease associations. WPSMDA is an alternative of MDA-CNN. The detail of WPSMDA can be found in the supplementary document. 10-fold cross validation is applied in the test. It is noted that CIPHER and PBMDA are two-layer-network-based models, therefore their input includes disease and miRNA layers as well as the miRNA-disease associations contained in the training set. For all tested methods, the performance is tested based on the testing set. Different metrics including Area Under the Receiver Operating Characteristic curve (AUROC), Area Under the Precision-Recall curve (AUPR), Precision, Recall and F1 score were used for performance evaluation.

Our results show that MDA-CNN achieves the highest performance in all tested methods. The average AUROC achieves by MDA-CNN across

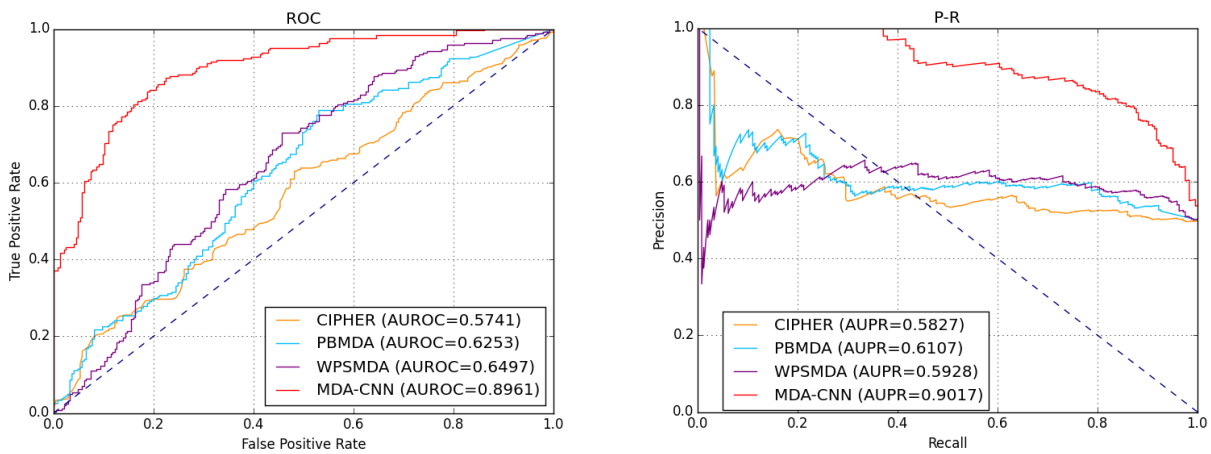

Fig. 3. The ROC (a) and P-R (b) curves of four tested methods on miRNA-disease association prediction task.

Table 1. The AUROC, AUPR, Precision, Recall and F1-Scores of four tested methods on miRNA-disease association prediction task.

	AUROC	AUPR	Precision	Recall	F1-score
CIPHER	0.5564	0.5612	0.4942	0.9954	0.6605
PBMDA	0.6321	0.6140	0.5192	0.9036	0.6594
WPSMDA	0.6406	0.5942	0.5418	0.9363	0.6864
MDA-CNN	0.8897	0.8887	0.8244	0.8056	0.8144

Table 2. The AUROC, AUPR, Precision, Recall and F1-Scores of four tested methods on miRNA-phenotype association prediction task.

	AUROC	AUPR	Precision	Recall	F1-score
CIPHER	0.5212	0.5271	0.5051	0.9991	0.6710
PBMDA	0.7453	0.7197	0.6725	0.7614	0.7141
WPSMDA	0.6927	0.6954	0.5492	0.9310	0.6909
MDA-CNN	0.9429	0.9344	0.8667	0.8748	0.8704

10-fold cross validation is 0.8897, which is significantly higher than the scores of CIPHER, PBMDA, WPSMDA (Table 1). Figure 3(a) shows ROC curves of one cross-validation result. The other nine results can be found in the supplementary document (Supplementary Figure S2). Comparing the AUPR scores from the four methods shows that MDA-CNN performs the best with the PBMDA as the runner-up (Table 1 and Figure 3(b)). The average AUPR scores of MDA-CNN, WPSMDA, PBMDA and CIPHER are 0.8887, 0.5942, 0.6140 and 0.5612 respectively. Figure 3(b) shows P-R curves of one cross-validation result. The other nine results can be found in the supplementary document (Supplementary Figure S3). We also listed the highest F1 scores that each method can be achieved in Table 1. The corresponding precision and recall scores are also listed. It is shown that MDA-CNN can achieve the highest F1 score. Importantly, both precision and recall scores of MDA-CNN are more than 0.80 when achieving the highest F1 score. In summary, this test shows that MDA-CNN can achieve significant improvement in predicting miRNA-disease associations.

3.4 Performance evaluation on predicting miRNA-phenotype associations

In addition to evaluate MDA-CNN on miRNA-disease association prediction problem, we further tested whether MDA-CNN can be applied to predict miRNA-phenotype associations. Patient phenotypes, which may be determined by both genetically or environmentally, are the physical, biochemical and physiological makeup of a patient (Robinson *et al.*, 2008). Understanding the miRNA-phenotype association can help people to reveal how miRNAs are implicated in human diseases.

Similar with the evaluation on miRNA-disease dataset, we compared MDA-CNN with three methods (i.e. CIPHER, PBMDA and WPSMDA) and applied 10-fold cross validation in the test. In general, the results show that MDA-CNN performs better than other methods, indicating that three layer-based feature representation model and proposed convolutional neural network model could increase the overall performance. The average AUROC score of MDA-CNN is 0.9429, which is significantly higher than the second best method PBMDA (the value is 0.7453) (Table 2). Similarly, MDA-CNN achieves the highest AUPR score (0.9344), while the scores of CIPHER, PBMDA and WPSMDA are 0.5271, 0.7197 and 0.6954 respectively (Table 2). Figure 4(a) and (b) shows ROC and PR curves of one cross-validation result. The other nine results can be found in supplementary document (Supplementary Figure S4 and S5). We also listed the highest F1 scores that each method can be achieved in Table 2. The corresponding precision and recall scores are also listed. It is shown that MDA-CNN can achieve the highest F1 score (0.8704). Comparing with other methods, the result shows that MDA-CNN can achieve both high precision and recall at the same time. In summary, MDA-CNN performs well not only on miRNA-disease association prediction but also miRNA-phenotype association prediction.

3.5 Effects of MDA-CNN components

In order to evaluate the performance of each step of MDA-CNN, we compared MDA-CNN with three versions of MDA-CNN, each with a different approach in feature representation step and convolutional neural network step on miRNA-disease associations dataset. To test the performance of our feature representation model, we created BR-CNN where the values in the feature vector (calculated based on the regression model) were replaced by the binary values based on the original input. To test the effect of the convolutional neural network model, we created MDA-SVM where the convolutional neural network model was replaced by Support Vector Machine (SVM) model.

Table 3 shows that MDA-CNN is better than the two versions, indicating that the two steps in MDA-CNN contribute to the performance and have been appropriately designed. The ROC curves and PR curves of one

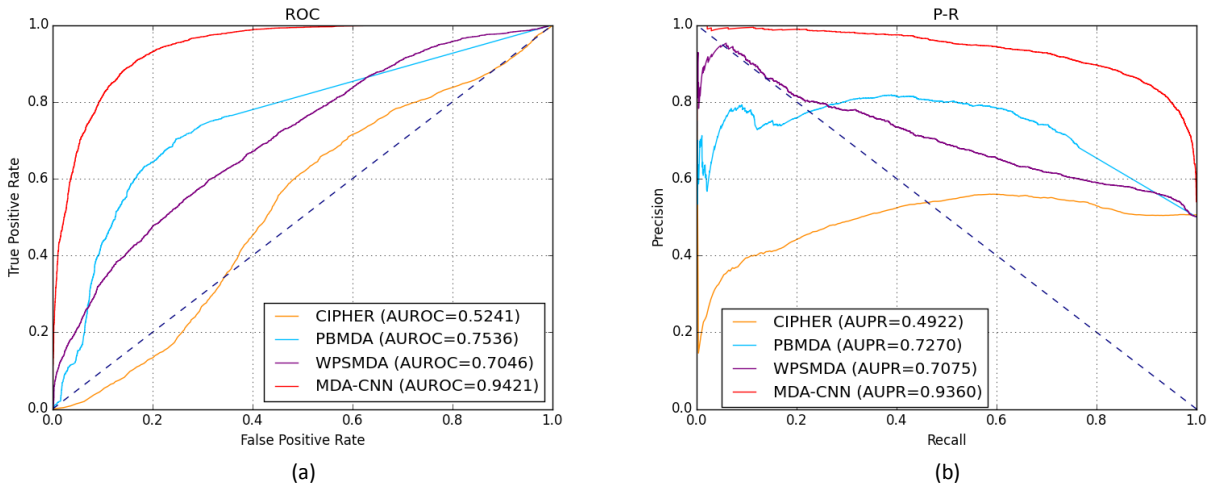

Fig. 4. The ROC (a) and P-R (b) curves of four tested methods on miRNA-phenotype association prediction task.

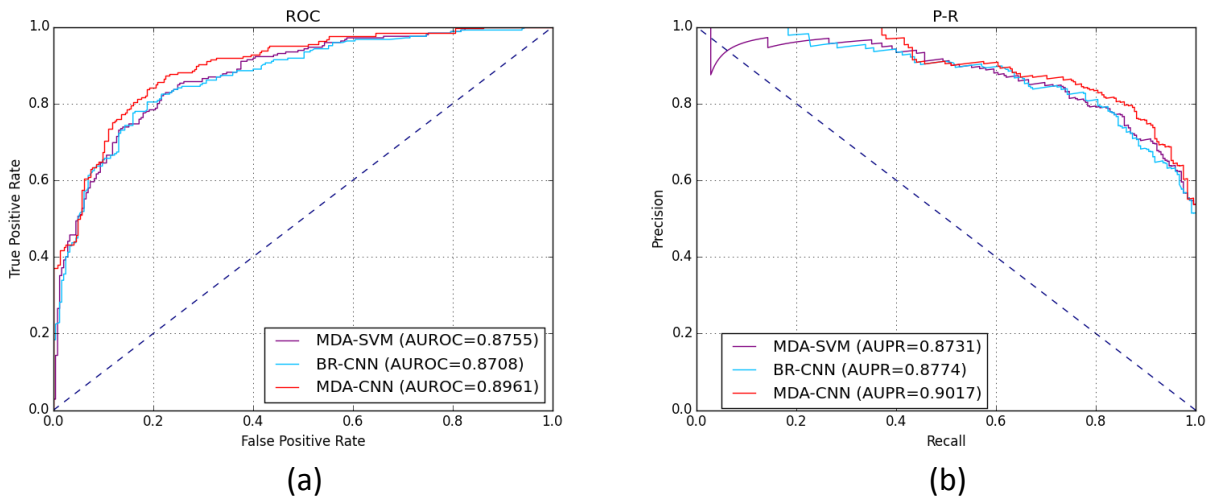

Fig. 5. The ROC (a) and P-R (b) curves of MDA-SVM, BR-CNN and MDA-CNN on miRNA-disease association prediction task.

Table 3. The AUROC, AUPR, Precision, Recall and F1-Scores of MDA-SVM, BR-CNN and MDA-CNN on miRNA-disease association prediction task.

	AUROC	AUPR	Precision	Recall	F1-score
MDA-SVM	0.8647	0.8524	0.7879	0.7856	0.7864
BR-CNN	0.8537	0.8515	0.7884	0.7726	0.7797
MDA-CNN	0.8897	0.8887	0.8244	0.8056	0.8144

cross-validation result are shown in Figure 5. The other nine results can be found in Supplementary document Figure S6 and S7.

3.6 Case studies on lung cancer and stomach cancer

We applied MDA-CNN to predict the miRNAs associated with lung cancer and stomach cancer.

Lung cancer is the leading cause of cancer-associated death (Bandi *et al.*, 2009; Jemal *et al.*, 2010). In the experiment, to avoid logic circle, we first removed all the associations between lung cancer and its related miRNAs from the training set. Then, we used MDA-CNN to predict the association between lung cancer and each miRNAs. The miRNAs are ranked based on their prediction probabilities associating with lung cancer. Based on the HMDD database, we found that 17 of top 20 miRNAs are associated with lung cancer (see supplementary table S1). The rest three miRNAs are hsa-mir-16, hsa-mir-15a and hsa-mir-106b, which are in the third, fourth and tenth place in the predicted miRNAs. Based on the literature study, we found that hsa-mir-15a and hsa-mir-16 were likely contributing to the tumorigenesis of non-small cell lung cancer by regulating cyclins D1, D2, and E1 (Bandi *et al.*, 2009). Although no direct evidence shows that hsa-mir-106b is associated with lung cancer, evidence shows that hsa-mir-106b is a cancer-associated miRNA (Nagini, 2012).

We also applied MDA-CNN to predict the miRNAs associated with stomach cancer, which is the second most common cause of cancer death in the world (Nagini, 2012; Siegel *et al.*, 2017). The experimental procedure was same as the aforementioned test on lung cancer. In our result, 19 of top 20 predicted associations between stomach cancer and miRNAs can be found in the HMDD database (see supplementary table S1). The only exception was hsa-mir-29b. Based on the literature study, we found that DNA methylation-associated silencing of miR-29b/c was critical for

stomach cancer development and contributed to the metastasis phenotype in stomach cancer (Cui *et al.*, 2015).

4 Conclusion

Recently, researchers have started to focus on identifying miRNA-disease associations by computational tools. In this article, we proposed a novel method named MDA-CNN to identify the miRNA-disease/phenotype associations. We first extract features of miRNA and disease/phenotype based on a three-layer network. Then, an auto-encoder-based model was proposed for feature selection. Using this feature representation, we proposed a convolutional neural network architecture for the purpose of predicting miRNA-disease/phenotype associations. To demonstrate the advantages of MDA-CNN, we compare it with three methods. The experiments on both miRNA-disease and miRNA-phenotype associations show that MDA-CNN performs better than existing methods. Furthermore, in order to evaluate the contribution of each step of MDA-CNN, we compared MDA-CNN with two versions of MDA-CNN, each with modifying the feature representation step or replacing convolutional neural network architecture with SVM. Overall, the evaluation test shows that both steps contribute to MDA-CNN performance and have been appropriately designed. Additionally, case studies on lung cancer and stomach cancer show that MDA-CNN can predict the miRNA-disease associations precisely.

Funding

This work has been supported by National Natural Science Foundation of China (Grant No. 61602386 and 61332014).

References

Ambros, V. (2003). Microrna pathways in flies and worms: growth, death, fat, stress, and timing. *Cell*, **113**(6), 673–676.

Angermueller, C., Pärnamaa, T., Parts, L., and Stegle, O. (2016). Deep learning for computational biology. *Molecular systems biology*, **12**(7), 878.

Ardekani, A. M. and Naeini, M. M. (2010). The role of micrornas in human diseases. *Avicenna journal of medical biotechnology*, **2**(4), 161.

Atwood, J. and Towsley, D. F. (2015). Diffusion-convolutional neural networks. *neural information processing systems*, pages 1993–2001.

Bandi, N., Zbinden, S., Gugger, M., Arnold, M., Kocher, V., Hasan, L., Kappeler, A., Brunner, T., and Vassella, E. (2009). mir-15a and mir-16 are implicated in cell cycle regulation in a rb-dependent manner and are frequently deleted or down-regulated in nonâŁ"small cell lung cancer. *Cancer Research*, **69**(13), 5553–5559.

Baolin, L. and Bo, H. (2007). Hprd: a high performance rdf database. *Network and Parallel Computing*, pages 364–374.

Boer, P. T. D., Kroese, D. P., Mannor, S., and Rubinstein, R. Y. (2005). A tutorial on the cross-entropy method. *Annals of Operations Research*, **134**(1), 19–67.

Bourlard, H. and Kamp, Y. (1988). Auto-association by multilayer perceptrons and singular value decomposition. *Biological Cybernetics*, **59**, 291–294.

Carleton, M., Cleary, M. A., and Linsley, P. S. (2007). Micrornas and cell cycle regulation. *Cell Cycle*, **6**(17), 2127.

Chen, X. and Yan, G. Y. (2014). Semi-supervised learning for potential human microrna-disease associations inference. *Scientific Reports*, **4**, 5501.

Chen, X., Liu, M. X., and Yan, G. Y. (2012). Rwrmda: predicting novel human microrna-disease associations. *Molecular Biosystems*, **8**(10), 2792.

Chen, X., Yan, C. C., Zhang, X., You, Z. H., Deng, L., Liu, Y., Zhang, Y., and Dai, Q. (2016). Wbsmda: Within and between score for mirna-disease association prediction. *Scientific Reports*, **6**, 21106.

Chicco, D., Sadowski, P. J., and Baldi, P. (2014). Deep autoencoder neural networks for gene ontology annotation predictions. pages 533–540.

Cui, H., Wang, L., Gong, P., Zhao, C., Zhang, S., Zhang, K., Zhou, R., Zhao, Z., and Fan, H. (2015). Deregulation between mir-29b/c and dnmt3a is associated with epigenetic silencing of the cdh1 gene, affecting cell migration and invasion in gastric cancer. *PloS one*, **10**(4), e0123926.

Dweep, H. and Gretz, N. (2015). mirwalk2. 0: a comprehensive atlas of microrna-target interactions. *Nature methods*, **12**(8), 697–697.

Goulart, L. F., Bettella, F., Sønderby, I. E., Schork, A. J., Thompson, W. K., Mattingsdal, M., Steen, V. M., Zuber, V., Wang, Y., Dale, A. M., *et al.* (2015). Micrornas enrichment in gwas of complex human phenotypes. *BMC genomics*, **16**(1), 304.

Grover, A. and Leskovec, J. (2016). node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864. ACM.

Han, K., Xuan, P., Ding, J., Zhao, Z. J., Hui, L., and Zhong, Y. L. (2014). Prediction of disease-related micrornas by incorporating functional similarity and common association information. *Genetics & Molecular Research Gmr*, **13**(1), 2009.

Hinton, G. E. and Salakhutdinov, R. (2006). Reducing the dimensionality of data with neural networks. *Science*, **313**(5786), 504–507.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *Computer Science*, **3**(4), pÃ¡gs. 212–223.

Hosmer, D. W. and Lemeshow, S. (2000). Applied logistic regression. *Technometrics*.

Huang, J., Wang, F., Argyris, E., Chen, K., Liang, Z., Tian, H., Huang, W., Squires, K., Verlinghieri, G., and Zhang, H. (2007). Cellular micrornas contribute to hiv-1 latency in resting primary cd4+ t lymphocytes. *Nature medicine*, **13**(10), 1241–1247.

Huang, Y., Shen, X. J., Zou, Q., Wang, S. P., Tang, S. M., and Zhang, G. Z. (2011). Biological functions of micrornas: a review. *Journal of physiology and biochemistry*, **67**(1), 129–139.

Jemal, A., Siegel, R., Ward, E., Murray, T., Xu, J., and Thun, M. J. (2010). Cancer statistics, 2007. *Ca A Cancer Journal for Clinicians*, **57**(1), 43–66.

Jiang, Q., Wang, Y., Hao, Y., Juan, L., Teng, M., Zhang, X., Li, M., Wang, G., and Liu, Y. (2008). mir2disease: a manually curated database for microrna deregulation in human disease. *Nucleic acids research*, **37**(suppl_1), D98–D104.

Jiang, Q., Hao, Y., Wang, G., Juan, L., Zhang, T., Teng, M., Liu, Y., and Wang, Y. (2010). Prioritization of disease micrornas through a human phenome-microrna-ome network. *BMC systems biology*, **4**(1), S2.

Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U., and Segal, E. (2007). The role of site accessibility in microrna target recognition. *Nature Genetics*, **39**(10), 1278–84.

Kim, J., Inoue, K., Ishii, J., Vanti, W. B., Voronov, S. V., Murchison, E., Hannon, G., and Abeliovich, A. (2007). A microrna feedback circuit in midbrain dopamine neurons. *Science*, **317**(5842), 1220–1224.

Kim, Y. (2014). Convolutional neural networks for sentence classification. *empirical methods in natural language processing*, pages 1746–1751.

Kingma, D. P. and Ba, J. L. (2015). Adam: A method for stochastic optimization. *international conference on learning representations*.

Kohavi, R. *et al.* (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. pages 1097–1105.

Lawrence, S., Giles, C. L., Tsoi, A. C., and Back, A. D. (1997). Face recognition: a convolutional neural-network approach. *IEEE Transactions on Neural Networks*, **8**(1), 98–113.

Lecun, Y., Boser, B. E., Denker, J. S., Henderson, D., Howard, R. E., and Hubbard, W. (1989). Backpropagation applied to handwritten zip code. *Neural Computation*.

Lecun, Y., Kavukcuoglu, K., and Farabet, C. (2010). Convolutional networks and applications in vision. pages 253–256.

Lecun, Y., Bengio, Y., and Hinton, G. E. (2015). Deep learning. *Nature*, **521**(7553), 436–444.

Lewis, B. P., Shih, I.-h., Jones-Rhoades, M. W., Bartel, D. P., and Burge, C. B. (2003). Prediction of mammalian microrna targets. *Cell*, **115**(7), 787–798.

Li, Y., Zhuang, L., Wang, Y., Hu, Y., Wu, Y., Wang, D., and Xu, J. (2013a). Connect the dots: a systems level approach for analyzing the mirna-mediated cell death network. *Autophagy*, **9**(3), 436–439.

Li, Y., Qiu, C., Tu, J., Geng, B., Yang, J., Jiang, T., and Cui, Q. (2013b). Hmdd v2. 0: a database for experimentally supported human microrna and disease associations. *Nucleic acids research*, **42**(D1), D1070–D1074.

Liu, B., Fang, L., Liu, F., Wang, X., Chen, J., and Chou, K.-C. (2015). Identification of real microrna precursors with a pseudo structure status composition approach. *PloS one*, **10**(3), e0121501.

Ma, L., Teruya-Feldstein, J., and Weinberg, R. A. (2007). Tumour invasion and metastasis initiated by microrna-10b in breast cancer. *Nature*, **449**(7163), 682–688.

Masino, A. J., Dechene, E. T., Dulik, M. C., Wilkens, A., Spinner, N. B., Krantz, I. D., Pennington, J. W., Robinson, P. N., and White, P. S. (2014). Clinical phenotype-based gene prioritization: an initial study using semantic similarity and the human phenotype ontology. *BMC bioinformatics*, **15**(1), 248.

Meister, G. and Tuschl, T. (2004). Mechanisms of gene silencing by double-stranded rna. *Nature*, **431**(7006), 343–349.

Min, S., Lee, B., and Yoon, S. (2017). Deep learning in bioinformatics. *Briefings in bioinformatics*, **18**(5), 851–869.

Nagini, S. (2012). Carcinoma of the stomach: A review of epidemiology, pathogenesis, molecular genetics and chemoprevention. *World Journal of Gastrointestinal Oncology*, **4**(7), 156.

Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. pages 807–814.

Pallez, D., Gardès, J., and Pasquier, C. (2017). Prediction of mirna-disease associations using an evolutionary tuned latent semantic analysis. *Scientific Reports*, **7**(1), 10548.

Ranzato, M., Huang, F. J., Boureau, Y., and Lecun, Y. (2007). Unsupervised learning of invariant feature hierarchies with applications to object recognition. pages 1–8.

Robinson, P. N., Köhler, S., Bauer, S., Seelow, D., Horn, D., and Mundlos, S. (2008). The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. *The American Journal of Human Genetics*, **83**(5), 610–615.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1988). Learning representations by back-propagating errors. *Nature*, **323**(6088), 533–536.

Shenoy, A. and Blelloch, R. H. (2014). Regulation of microrna function in somatic stem cell proliferation and differentiation. *Nature Reviews Molecular Cell Biology*, **15**(9), 565–576.

Shi, H., Xu, J., Zhang, G., Xu, L., Li, C., Wang, L., Zhao, Z., Jiang, W., Guo, Z., and Li, X. (2013). Walking the interactome to identify human mirna-disease associations through the functional link between mirna targets and disease genes. *BMC systems biology*, **7**(1), 101.

Siegel, R. L., Miller, K. D., and Jemal, A. (2017). Cancer statistics, 2017. *CA: A Cancer Journal for Clinicians*, **67**(1), 5.

Taganov, K. D., Boldin, M. P., Chang, K. J., and Baltimore, D. (2006). Nf-Îºb-dependent induction of microrna mir-146, an inhibitor targeted to signaling proteins of innate immune responses. *Proceedings of the National Academy of Sciences of the United States of America*, **103**(33), 12481.

Volinia, S., Galasso, M., Sana, M. E., Wise, T. F., Palatini, J., Huebner, K., and Croce, C. M. (2012). Breast cancer signatures for invasiveness and prognosis defined by deep sequencing of microrna. *Proceedings of the National Academy of Sciences*, **109**(8), 3024–3029.

Wax, M. and Ziv, J. (1977). Improved bounds on the local mean-square error and the bias of parameter estimators (corresp.). *IEEE Transactions on Information Theory*, **23**(4), 529–530.

Wei, L., Huang, Y., Qu, Y., Jiang, Y., and Zou, Q. (2012). Computational analysis of mirna target identification. *Current Bioinformatics*, **7**(4), 512–525.

Wu, X., Jiang, R., Zhang, M. Q., and Li, S. (2008). Network-based global inference of human disease genes. *Molecular systems biology*, **4**(1), 189.

Xuan, P., Han, K., Guo, M., Guo, Y., Li, J., Ding, J., Liu, Y., Dai, Q., Li, J., Teng, Z., *et al.* (2013). Prediction of micrornas associated with human diseases based on weighted k most similar neighbors. *PloS one*, **8**(8), e70204.

You, Z.-H., Huang, Z.-A., Zhu, Z., Yan, G.-Y., Li, Z.-W., Wen, Z., and Chen, X. (2017). Pbmda: A novel and effective path-based computational model for mirna-disease association prediction. *PLoS computational biology*, **13**(3), e1005455.

Yuan, Y., Liu, B., Xie, P., Zhang, M. Q., Li, Y., Xie, Z., and Wang, X. (2015). Model-guided quantitative analysis of microrna-mediated regulation on competing endogenous rnas using a synthetic gene circuit. *Proceedings of the National Academy of Sciences*, **112**(10), 3158–3163.

Zou, Q., Li, J., Song, L., Zeng, X., and Wang, G. (2015). Similarity computation strategies in the microrna-disease network: a survey. *Briefings in functional genomics*, **15**(1), 55–64.