DMFLDA: A deep learning framework for predicting lncRNA–disease associations

Min Zeng, Chengqian Lu, Zhihui Fei, Fang-Xiang Wu, Yaohang Li, Jianxin Wang and Min Li*

Abstract— A growing amount of evidence suggests that long non-coding RNAs (lncRNAs) play important roles in the regulation of biological processes in many human diseases. However, the number of experimentally verified lncRNA-disease associations is very limited. Thus, various computational approaches are proposed to predict lncRNA-disease associations. Current matrix factorization-based methods cannot capture the complex non-linear relationship between lncRNAs and diseases, and traditional machine learning-based methods are not sufficiently powerful to learn the representation of lncRNAs and diseases. Considering these limitations in existing computational methods, we propose a deep matrix factorization model to predict lncRNA-disease associations (DMFLDA in short). DMFLDA uses a cascade of non-linear hidden layers to learn latent representation to represent lncRNAs and diseases. By using non-linear hidden layers, DMFLDA captures the more complex non-linear relationship between lncRNAs and diseases than traditional matrix factorization-based methods. In addition, DMFLDA learns features directly from the lncRNA-disease interaction matrix and thus can obtain more accurate representation learning for lncRNAs and diseases than traditional machine learning methods. The low dimensional representations of the lncRNAs and diseases are fused to estimate the new interaction value. To evaluate the performance of DMFLDA, we perform leave-one-out cross-validation and 5-fold cross-validation on known experimentally verified lncRNA–disease associations. The experimental results show that DMFLDA performs better than the existing methods. The case studies show that many predicted interactions of colorectal cancer, prostate cancer, and renal cancer have been verified by recent biomedical literature.

Index Terms—Deep learning, matrix factorization, lncRNA-disease associations, non-linear features.

1 Introduction

With the rapid development of sequencing technology, researchers have discovered a fact that more than 98% of the human genome does not encode protein sequences [1]. Further studies indicate that lots of non-coding RNAs (ncRNAs) play critical roles in various fundamental and important biological processes [2-7]. NcRNAs can be divided into small ncRNA and long ncRNA based on the length of nucleotides [8]. Long non-coding RNAs (lncRNAs) have more than two hundred nucleotides and are a very important class of ncRNAs [9-11]. More and more evidence indicates that lncRNAs have very close associations with many human diseases such as breast cancer, lung cancer, and Alzheimer's disease. For example, lncRNA 'H19' not only has great effects on primary breast carcinomas [12, 13] but also confirmed to be associated with lung cancer [14]. The expression of lncRNA 'BACE1-AS' drives rapid feed-forward regulation of b-secretase in Alzheimer's disease [15]. Thus identification of potential lncRNA–disease associations is of great significance in biology, which can help understand the disease mechanism at the lncRNA level.

Considering the huge cost of traditional biological experiments, various computational methods have been developed to predict potential lncRNA–disease associations. These methods can be divided into three categories. The first category is based on machine learning methods. Chen et al. constructed a semi-supervised learning framework called LRLSLDA to predict potential disease-related lncRNAs [16]. LRLSLDA integrates the known disease–lncRNA associations and lncRNA expression profiles and utilizes Laplacian regularized least squares to optimize the objective function. Zhao et al. applied a naive Bayesian classifier to predict cancer-related lncRNAs by integrating genomic, miRNA targets and expression features [17]. Lan et al. used a bagging SVM to predict potential lncRNA-disease associations by fusing lncRNA similarity and disease similarity [18]. Fu et al. proposed a matrix factorization based model named MFLDA to predict lncRNA-disease associations [19]. Lu et al. used an inductive matrix completion model called SIMCLDA to predict lncRNA-disease associations [20]. The second category is based on biological networks. Sun et al. constructed a lncRNA–lncRNA functional similarity network and used the random walk technique to predict lncRNA–disease associations [21]. Yao et al. used a multi-level composite network to prioritize candidate lncRNAs associated with diseases by integrating genes, lncRNAs, phenotypes, and interactions [22]. Chen et al. applied the model of KATZ measure to predict lncRNA-disease associations by integrating known lncRNA-disease associations, lncRNA expression profiles, lncRNA functional similarity, disease semantic similarity, and Gaussian in-

- *M. Zeng, C. Lu, Z. Fei, J. Wang, and M. Li are with the School of Computer Science and Engineering, Central South University, Changsha, P.R. China. Email: {zengmin, chengqlu}@csu.edu.cn, zhihuifei@foxmail.com, {jxwang, limin}@mail.csu.edu.cn.*
- *F.X.Wu is with the Division of Biomedical Engineering and Department of Mechanical Engineering, University of Saskatchewan, Saskatoon, SKS7N5A9, Canada. E-mail: faw341@mail.usask.ca.*
- *Y. Li is with the Department of Computer Science, Old Dominion University, Norfolk, VA23529, USA. Email: yaohang@cs.odu.edu*
- ** Corresponding author*

teraction profile kernel similarity [23]. Zhou et al. integrated three networks (lncRNA–lncRNA crosstalk network, disease–disease similarity network and known lncRNA–disease association network) into a heterogeneous network and applied the random walk technique to predict lncRNA-disease associations [24]. Zhang et al. used a flow propagation algorithm to integrate multiple networks based on lncRNA similarity, protein-protein interactions, disease similarity, and the associations to predict lncRNA-disease associations [25]. The third category is not based on known lncRNA–disease associations. The above two types of methods used the known lncRNA–disease associations. However, known experimentally verified associations are quite small. Thus some researchers used the other biological information to predict lncRNA–disease associations. Liu et al. combined human lncRNA expression profiles, gene expression profiles, and human disease-associated gene data to propose a computational framework [26]. Chen integrated miRNA-disease associations and lncRNA-miRNA interactions to predict lncRNA-disease associations [27].

In recent years, deep learning methods have been successfully applied in various fields including computer vision, natural language processing and bioinformatics [28-35]. Inspired by their success, we propose a deep matrix factorization framework called DMFLDA to automatically predict lncRNA-disease associations only using lncRNA-disease interaction matrix. The basic idea of deep matrix factorization model [36, 37] is to treat prediction of lncRNA-disease associations as a recommendation system problem and use a cascade of non-linear hidden layers to learn a latent low dimensional space to represent the lncRNAs and diseases. Then the low dimensional representations of lncRNAs and diseases are fused to estimate new interaction values. There are two advantages of using a deep matrix factorization model. First, compared with traditional matrix factorization-based methods, deep matrix factorization model can learn the non-linear, more complex relationships between lncRNAs and diseases. As we know, traditional matrix factorization methods can only capture the linear relationship between lncRNAs and diseases. However, the relationship between lncRNAs and diseases are too complicated, such complex relationships can be difficult to characterize with linear models. Second, compared with traditional machine learning-based methods (support vector machine, naïve Bayes, etc.), the deep matrix factorization model is more powerful to represent the features of lncRNAs and diseases. Feature extraction directly from the lncRNA-disease interaction matrix can obtain more accurate representation learning for lncRNAs and diseases than traditional machine learning methods. Although traditional machine learning methods have obtained good results, there is room for improvement. A powerful deep learning-based model can help us better predict lncRNA-disease interactions.

In this study, we only use lncRNA-disease interactions to construct a deep matrix factorization model. A lot of biological information, such as disease similarity, lncRNA similarity, and lncRNA expression profiles, are useful for the prediction of lncRNA-disease interactions, while a lot of lncRNAs do not have such information. In addition, there are many different methods for measuring disease and lncRNA similarity, such as Pearson, Spearman, and Jaccard similarity coefficient. The commonly used method for the similarity coefficient selection is based on the results of statistical methods and machine learning methods. As a result, it is difficult to explain why this similarity coefficient was used instead of that one. We thus focus on lncRNA-disease interactions, which are wildly used in various studies.

To evaluate the performance of our model, we carry out extensive experiments on a preprocessed gold standard dataset. The experimental results show that DMFLDA reaches the AUC values of 0.8393 and 0.8288 in leave-one-out cross-validation and 5-fold cross-validation, respectively, which outperform other computational methods including SIMCLDA [20], MFLDA [19], TPGLDA [38], and LDAP [18]. In order to further evaluate the capability of DMFLDA, we conduct case studies for three types of cancers including colorectal cancer, prostate cancer, and renal cancer. Case studies show that 22 out of 30 (8 for colorectal cancer, 7 for prostate cancer and 7 for renal cancer) cancer-related lncRNAs predicted by DMFLDA are verified by manually mining recent biomedical literature. These findings show the capability of DMFLDA for predicting potential lncRNA–disease associations.

2 METHODS

2.1 Problem definition

Given M lncRNAs $R = \{r_1, r_2, r_3, \ldots, r_m\}$, N diseases $D = \{d_1, d_2, d_3, \ldots, d_n\}$. We define the lncRNA–disease interaction matrix by $R \in R^{m \times n}$ based on known lncRNA–disease association datasets. We construct the interaction matrix as follows.

$$R_{ij} = \begin{cases} 1, & if\ lncRNA\ i\ is\ linked\ to\ disease\ j \\ 0, & if\ the\ relationship\ is\ unknown \end{cases} \quad (1)$$

The purpose of lncRNA–disease association prediction is predicting the score of unknown relationship in interaction matrix R by using known lncRNA–disease associations. There are a lot of zero entries in this interaction matrix R, the true interactions we already know only accounts for a small percentage. The sparsity of interaction information brings great challenges to our prediction.

2.2 Traditional matrix factorization

Given a partially observed matrix $R \in R^{m \times n}$, U and V denote two matrices which have dimensionality of $m \times d$ and $n \times d$, respectively. In general, matrix R is sparse and d is much smaller than m and n. The crucial step of traditional matrix factorization is the decomposition of the partially observed matrix R into matrices U and V. Then the product UV^T can closely reconstruct the interaction matrix R. A new estimate at the *(i, j)* position of interaction matrix R is predicted as follows:

$$R_{ij}^{pred} = u_i v_j^T \quad (2)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2020.2983958, IEEE/ACM Transactions on Computational Biology and Bioinformatics

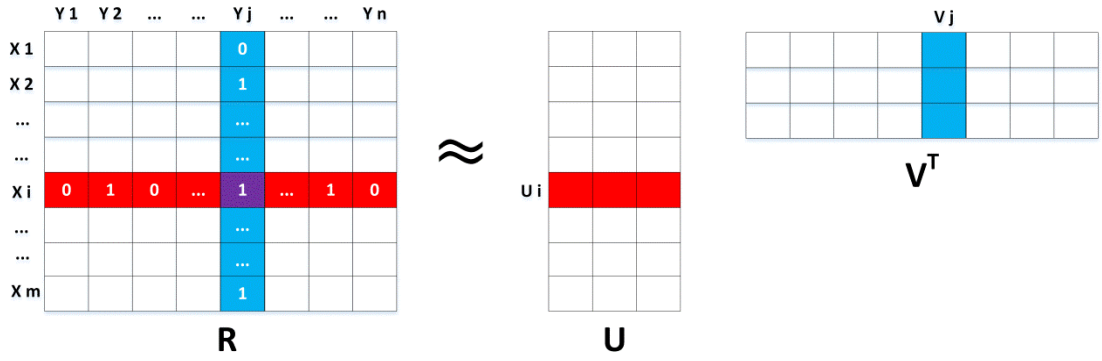


Figure 1. Illustration of the decomposition of matrix *R* into matrices *U* and V^T.

where R_{ij}^{pred} denotes a new estimate, u_i and v_j represent the latent factor of U and V, respectively. Figure 1 gives the illustration of matrix factorization. The solution of matrix factorization is obtained by minimizing the regularized loss function on the data:

$$Loss = \sum_{i,j\in\Omega}(R_{ij} - u_i^T v_j)^2 + \lambda(|u_i|^2 + |v_j|^2) \quad (3)$$

where Ω denotes the known observed interactions of R, the constant λ is a parameter to adjust the weight between empirical risk and the regularized term.

2.3 Deep Matrix Factorization Model

In our previous study [20], we formulated lncRNA–disease association prediction as a recommendation system problem. In recent years, deep learning technology has been successfully and widely used in various fields, such as image classification, natural language processing, and speech recognition. In addition, some researchers began to explore how to apply the deep learning techniques to the matrix factorization framework. Inspired by their work, we construct a deep matrix factorization model to predict lncRNA–disease associations. Figure 2 gives a schematic view of our deep matrix factorization model.

In our deep matrix factorization model, each lncRNA is represented as a row of the interaction matrix, which describes the lncRNA's interaction across all diseases. Each disease is represented as a column of the interaction matrix, which describes the disease's interaction across all lncRNAs. The rows and columns of the interaction matrix are considered to be lncRNA's and disease's feature vector. The input layer of the model consists of two feature vectors, lncRNA and disease feature vectors. Here we need to mention a detail. In the interaction matrix, the intersection of a row and a column are considered the label of a sample in supervised learning. Thus we cannot directly use the label in the feature vector to represent lncRNAs and diseases. To solve the problem, we mask the value of the intersection and replace it with 0. Then the two feature vectors are fed into two fully connected layers that project the sparse representation to dense representation. The obtained lncRNA and disease vectors can be seen as the latent vectors for a lncRNA and a disease. Then we fuse a pair of lncRNA and disease latent vectors to a new vector which is fed into a fully connected layer with a sigmoid activation function to perform prediction task. Formally, x and y denote the lncRNA and disease input vectors, respectively. $W_{x1}, W_{x2}, W_{y1}, W_{y2}$ are weight matrices and $b_{x1}, b_{x2}, b_{y1}, b_{y2}$ are bias terms of each intermediate fully connected layer, $O_{x1}, O_{x2}, O_{y1}, O_{y2}$ represent the outputs of each intermediate fully connected layer. The ReLU function is used as the activation function of each intermediate fully connected layer.

$$ReLU(x) = max(0, x) \quad (4)$$

The outputs of the first fully connected layer are:

$$O_{x1} = ReLU(W_{x1}x + b_{x1}) \quad (5)$$

$$O_{y1} = ReLU(W_{y1}y + b_{y1}) \quad (6)$$

Our deep matrix factorization model has two fully connected layers that transform the raw representation of lncRNAs and diseases to the latent dense representation of lncRNAs and diseases. Through the neural network, the latent dense vectors of lncRNAs and diseases are:

$$O_{x2} = ReLU(W_{x2}ReLU(W_{x1}x + b_{x1}) + b_{x2}) \quad (7)$$

$$O_{y2} = ReLU(W_{y2}ReLU(W_{y1}y + b_{y1}) + b_{y2}) \quad (8)$$

Then we use element-wise multiplication to fuse the latent dense vectors of a lncRNA and a disease into a new vector:

$$vec_{new} = Element - wise\ multiplication\ (O_{x2}, O_{y2}) \quad (9)$$

Element-wise multiplication is a standard operation which on each element of vectors while doing multiplication. For example, if V_1 = [1 2 3] and V_2 = [5 6 2], then multiply two vectors by multiplying all of the corresponding elements. The result w is:

$$w = Element - wise\ multiplication\ (V_1, V_2) = [5\ \ 12\ \ 6] \quad (10)$$

The new vector is fed into a fully connected layer with a sigmoid activation function to perform the prediction task.

$$sigmoid(x) = \frac{1}{1+exp\ (-x)} \quad (11)$$

$$R_{pred} = sigmoid(vec_{new}) \quad (12)$$

To learn the parameters in our deep matrix factorization model, we use a binary cross-entropy loss function as the loss function. The binary cross-entropy loss function is the most widely used loss function in classification. Its optimization can be done by performing adaptive moment estimation (Adam).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2020.2983958, IEEE/ACM Transactions on Computational Biology and Bioinformatics

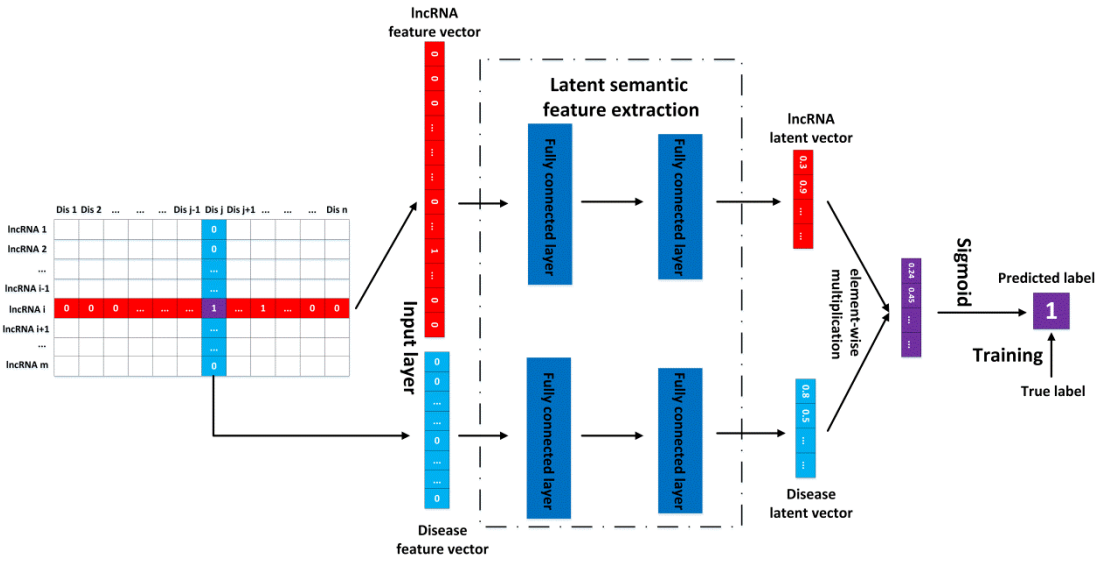


Figure 2. Illustration of our proposed deep matrix factorization model for prediction of lncRNA–disease associations. The red vector in interaction matrix is used to indicate lncRNA i's vector and the blue vector in interaction matrix is used to indicate disease j's vector. The two vectors are fed into two different networks to extract latent feature of them. The element-wise multiplication is applied to fuse the outputs of the two different networks and the result is represented by the purple vector. A sigmoid activation function is applied to the purple vector for classification.

$$Loss = \sum [R\, log(R_{pred}) + (1 - R)\, log(1 - R_{pred})] + \lambda(|\theta\,|\text{^}2) \qquad (13)$$

where θ is the weight vector, the regularization parameter λ is a parameter to adjust the weight between empirical risk and the regularized term.

It should be noted that the prediction of lncRNA–disease associations by using a deep matrix factorization model is a classification problem. As previously mentioned, the lncRNA–disease interaction matrix has two types of values, 1 and 0. We can view the value of 1 as a label of positive instance, and 0 as a label of negative instance. This matrix is very sparse for it has a lot of values of 0 and a few of values of 1. If we use all values of 0 for training, then the prediction of lncRNA–disease associations will become an extremely imbalanced learning problem due to the fact that the number of unobserved instances is far more than the number of known instances [39]. Thus we need to sample some negative instances from unobserved data for training. We apply a sampling method to our model and the basic idea of the sampling method is to use a balanced set of positive and negative instances to update parameters of the model in each epoch [40, 41]. Such a sampling method ensures that our model does not bias toward the positive or negative instances in each training epoch. We denote the set of known interactions by S^+, the set of all unobserved interactions by S^-. S^-_{samp} denotes the subset of negative instances which are sampled from S^- in each epoch. We keep the number of random chosen negative instances is equal to the number of positive instances in S^+. In each epoch, $S^+ \cup S^-_{samp}$ constitute the training dataset and we use the training dataset to update the parameters on our model.

3 Experimental Results

3.1 Data sources

In this study, we retrieve known lncRNA-disease associations from LncRNADisease [42], GeneRIF [43] and Lnc2Cancer [44]. After checking names of lncRNAs (according to Lncipedia, lncrnadb, HGNC, and NCBI) and diseases (according to Mesh, UMLS and NCBI), we remove all the repeating records and all the entries of other organisms. The statistics of the processed dataset are shown in Table 1.

TABLE 1.
STATISTICS OF THE DATASET WE USED IN THE STUDY.

Dataset	# of lncRNAs	# of diseases	# of interactions	Interaction density
	577	272	1583	1.008%

3.2 Evaluation metrics

To evaluate the performance of deep matrix factorization model and other methods in predicting lncRNA–disease associations, leave-one-out cross-validation (LOOCV) and 5-fold cross-validation (5-fold CV) are applied to our study as previously used in other studies. We use LOOCV on the known lncRNA–disease associations. Specifically: in each turn, for a given disease j_i, each known lncRNA associated to j_i is chosen as the test sample, and the other known lncRNA–disease associations and the same number of random sampled unknown lncRNA–disease associations are combined as the training dataset. A 5-fold CV is used to evaluate the capability of a model to predict potential associations. In the 5-fold CV, at each

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2020.2983958, IEEE/ACM Transactions on Computational Biology and Bioinformatics

turn, the known lncRNA-disease associations are divided into five subsets, four of which are used for training and the remaining one is used for testing.

After the training step, we can use our model to obtain all the new values of the interaction matrix. Then we calculate the true positive rate (TPR) and false positive rate (FPR):

$$TPR = \frac{TP}{TP+FN} \tag{14}$$

where TP represents the number of positive samples identified correctly and FN denotes the number of positive samples identified incorrectly.

$$FPR = \frac{FP}{FP+TN} \tag{15}$$

where FP represents the number of negative samples identified incorrectly and TN denotes the number of negative samples identified correctly.

The receiver operating characteristic (ROC) curve is drawn to illustrate the performances of models. It plots the TPR against the FPR at various threshold settings. AUC is defined as the area under the ROC curve. In general, a classifier that provides a larger AUC shows it has better performance.

3.3 Implementation details

We implement DMFLDA in Tensorflow which is a publicly available deep learning library developed by Google [45]. We use a row of the interaction matrix as an input vector of a lncRNA and a column of the interaction matrix as an input vector of a disease. For the intersection of a row and a column, we mask the value and replace it with 0. We use two fully connected layers to extract the latent features of lncRNAs and diseases. The number of neurons in the first fully connected layer is 48 and in the second layer is 32. Thus the dimension of latent vector of the lncRNAs and diseases is 32. We use element-wise multiplication to fuse them to a new vector which is also a 32-dimensional vector. The non-linear activation function is the ReLU function. Dropout rate of 0.05 is used on each fully connected layer in the network to avoid overfitting. The regularization parameter λ in loss function is set to 0.001. Weights in DMFLDA are initialized using the Gaussian distribution (with a mean of 0 and standard deviation of 0.01). The Adam optimizer is applied to train DMFLDA, the initial learning rate is set to 0.0005, and the batch size is set to 32. In the training process, we randomly sample one known lncRNA-diseases interactions as the validation data to evaluate the performance of DMFLDA.

3.4 Comparison with other methods

3.4.1 Compared methods

To evaluate the performance of DMFLDA in predicting lncRNA–disease associations, we compare DMFLDA with 4 popular computational methods (SIMCLDA, MFLDA, TPGLDA, and LDAP). SIMCLDA was developed by Lu et al., and based on inductive matrix completion [20]. SIMCLDA finds a low-rank matrix that can integrate prior knowledge of lncRNAs and diseases to complete the lncRNA–disease association matrix. MFLDA was developed by Fu et al., which decomposes data matrices of heterogeneous data sources into low-rank matrices via matrix tri-factorization to explore and exploit their intrinsic and shared structure [19]. TPGLDA was developed by Ding et al., which integrates gene-disease associations with lncRNA-disease associations and uses an effective resource allocation algorithm to predict potential lncRNA-disease associations [38]. LDAP was developed by Lan et al., which uses a bagging SVM to predict potential lncRNA-disease associations by fusing lncRNA similarity and disease similarity [18].

3.4.2 Results

We compared DMFLDA with 4 popular computational methods by using two kinds of evaluation methods (LOOCV and 5-fold CV). Figure 3a shows the ROC curves of DMFLDA and other compared methods in the LOOCV.

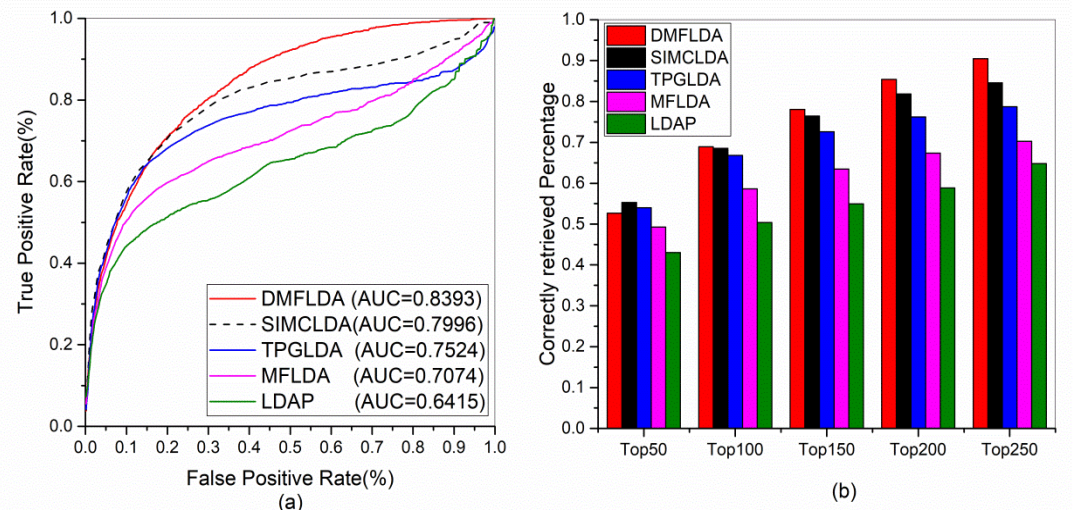


Figure 3. Comparison of performances obtained by DMFLDA and other computational methods in the LOOCV. (a) The ROC curves of DMFLDA and other compared methods. (b) Ratios of correctly retrieved known lncRNA–disease associations for specified rank thresholds.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2020.2983958, IEEE/ACM Transactions on Computational Biology and Bioinformatics

We can see that the ROC curve of DMFLDA is significantly higher than other methods. The AUC obtained by DMFLDA is 0.8393, which is better than SIMCLDA (0.7996), TPGLDA (0.7524), MFLDA (0.7074) and LDAP (0.6415). Our results indicate that our method exhibits extremely high accuracy. Furthermore, we evaluate the numbers of correctly retrieved lncRNA–disease associations. Specifically, each predicted association has a responding rank, if the rank higher than a specified rank threshold k, and then we regard it as a correctly retrieved association. Figure 3b shows the histograms of DMFLDA and other competing methods in the LOOCV. In top 50 (k=50), DMFLDA is slightly lower than SIMCLDA and TPGLDA, but higher than MFLDA and LDAP. In top 100, 150, and 200, we can see that DMFLDA can find more correct associations than other methods. Considering that we use less biological information than other methods, DMFLDA obtains good results actually. Figure 4a shows the ROC curves of DMFLDA and other compared methods in the 5-fold CV. Figure 4b shows the histograms of DMFLDA and other competing methods in the 5-fold CV. From Figure 4a, we can see that the ROC curve of DMFLDA is significantly higher than other methods. From Figure 4b, we can see that DMFLDA can find more correct associations than other methods. Overall, DMFLDA outperforms the other computational methods in the LOOCV and 5-fold CV.

3.5 The effects of hyper-parameters

In our model, some hyper-parameters, such as the number of neurons in each fully connected layer and the regularization parameter λ in the loss function, have different effects on experimental performance. To obtain the best performance of DMFLDA, we have tried a set of different hyper-parameters of network architectures to find the best hyper-parameters for predicting lncRNA-disease interactions.

3.5.1 The effects of the number of neurons in the last fully connected layer

The number of neurons in each fully connected layer is a sensitive parameter in our model, especially the last fully connected layer. The number of neurons in the last fully connected layer represents the dimension of the latent vector of lncRNAs and diseases and is very important for predicting lncRNA-disease interactions. However, it takes a lot of time to do experiments of different combinations of the hyper-parameters with LOOCV. For simplicity, we only compare the performance with different numbers of neurons in the last fully connected layer. The number of neurons in the first layer is set to 48. We change the number of neurons in the last layer from 8 to 48 (8, 16, 32 and 48) to find the best parameter. The results are shown in Table 2, the last layer with 32 neurons obtains the best performance.

TABLE 2.
AUC FOR MODELS WITH DIFFERENT NEURONS OF THE LAST FULLY CONNECTED LAYER.

# of neurons	8	16	32	48
AUC	0.8173	0.8310	0.8393	0.8385

TABLE 3.
AUC FOR MODELS WITH DIFFERENT PARAMETER.

λ	0.0001	0.0003	0.001	0.003	0.01	0.03
AUC	0.8317	0.8389	0.8393	0.8313	0.8278	0.8211

3.5.2 The effects of λ

The regularization parameter λ in the loss function is a trade-off parameter to adjust the weight between the empirical risk and the regularized term. In the loss function, we want to fit the training data well by using the empirical risk and keep the parameters of the deep learning

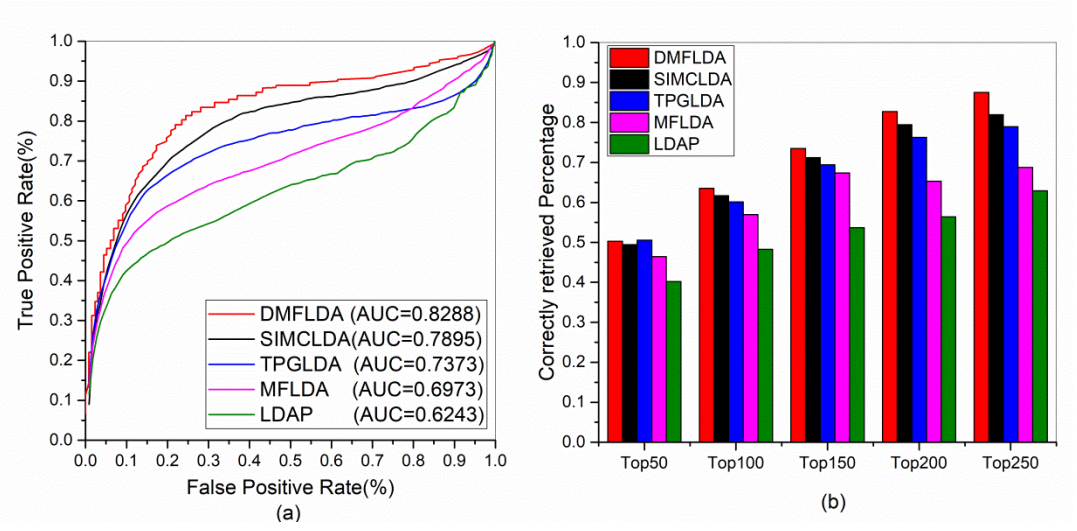


Figure 4. Comparison of performances obtained by DMFLDA and other computational methods in the 5-fold CV. (a) The ROC curves of DMFLDA and other compared methods. (b) Ratios of correctly retrieved known lncRNA–disease associations for specified rank thresholds.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2020.2983958, IEEE/ACM Transactions on Computational Biology and Bioinformatics

TABLE 4.
DMFLDA PREDICTED LNCRNAS ASSOCIATED WITH COLORECTAL CANCER (TOP 10) WITH THE CORRESPONDING REFERENCES.

lncRNA	Reference	Rank	Description
SPRY4-IT1	Cao et al. (2016)	1	Long noncoding rna SPRY4-IT1 promotes malignant development of colorectal cancer by targeting epithelial–mesenchymal transition.
CDKN2B-AS1	Chen et al. (2016)	2	See Table 1 in Ref Chen et al. (2016).
GAS5	Li et al. (2017)	3	Long non-coding RNA GAS5 acts as a tumour suppressor in colorectal cancer by inhibiting interleukin-10 and vascular endothelial growth factor expression.
KCNQ1OT1	Zhang et al. (2018)	4	While Sunamura et al. have demonstrated that beta-catenin can bind to KCNQ1OT1 promotor and activate its transcription in colon cancer cells.
BANCR	Su et al. (2015)	5	LncRNA BANCR is abnormally expressed in non-small cell lung cancer, melanoma, and colorectal cancer compared.
SNHG16	Christensen et al. (2016)	6	SNHG16 is regulated by the Wnt pathway in colorectal cancer and affects genes involved in lipid metabolism.
HULC	Dong et al. (2019)	7	Long non-coding RNA HULC interacts with miR-613 to regulate colon cancer growth and metastasis through targeting RTKN.
Sox4	Lin et al. (2013)	8	Clinical and prognostic implications of transcription factor SOX4 in patients with colon cancer.
IGF2-AS	Unknown	9	N/A
NBAT1	Unknown	10	N/A

TABLE 5.
DMFLDA PREDICTED LNCRNAS ASSOCIATED WITH PROSTATE CANCER (TOP 10) WITH THE CORRESPONDING REFERENCES.

lncRNA	Reference	Rank	Description
UHRF1	Jazirehi et al. (2012)	1	UHRF1: a master regulator in prostate cancer.
PANDAR	Unknown	2	N/A
CCAT1	Mizrahi et al. (2015)	3	CCAT1 is a 2628 nucleotide lncRNA located on chromosome 8q24.21 in an intergenic area described before as a “hot spot” harboring multiple genetic alternations in both colon and prostate cancer.
RN7SK	Unknown	4	N/A
XIST	Du et al. (2017)	5	LncRNA XIST acts as a tumor suppressor in prostate cancer through sponging miR-23a to modulate RKIP expression.
TINCR	Dong et al. (2018)	6	LncRNA TINCR is associated with clinical progression and serves as tumor suppressive role in prostate cancer.
LINC-ROR	Liu et al. (2017)	7	Curcumin suppresses proliferation and in vitro invasion of human prostate cancer stem cells by ceRNA effect of miR-145 and lncRNA-ROR.
NPTN-IT1	Unknown	8	N/A
TDRG1	Wang et al. (2016)	9	The knockout of TDRG1 significantly decreased the phosphorylation levels of PI3K/p85, PI3K/p110, Akt, and mammalian target of rapamycin. The PI3K/Akt/mTOR pathway plays an important role in cell growth and survival. Similar findings have been reported in ovarian, colorectal, and prostate cancers.
CRNDE	Ellis et al. (2012)	10	While CRNDE expression is mildly elevated in prostate cancer.

model small by using the regularized term. Parameter λ controls the trade-off between the two terms and is a sensitive parameter in our model. The number of neurons in each fully connected layer is set to 48 and 32, respectively. We train our model with the different parameters of 0.0001, 0.0003, 0.001, 0.003, 0.01 and 0.03 for λ to find the best parameter with LOOCV. We show our results in Table 3 and find the best performance is obtained when λ is 0.001.

3.6 Case studies

In order to further evaluate the capability of DMFLDA, we conduct case studies for three kinds of important cancers including colorectal cancer, prostate cancer, and re-

nal cancer. For a target cancer, we make use of the already trained model to estimate new values for those lncRNAs that do not have the interactions with the target cancer. Then we select the top 10 plausible lncRNAs as our predicted lncRNAs for the target cancer. After that, we check the top 10 plausible lncRNAs by manually mining recent biomedical literature.

Colorectal cancer, also known as colon cancer, is a great threat to public health, being the third most commonly diagnosed cancer in males and the second in females worldwide [46]. More than 1 million new colorectal cancer cases were diagnosed each year since 2012. It is very important to find the associations between colorectal cancer and some lncRNAs. DMFLDA is applied to infer potential colorectal cancer-related lncRNAs. As a result, 8 colorectal cancer-related lncRNAs (SPRY4-IT1, CDKN2B-AS1, GAS5, KCNQ1OT1, BANCR, SNHG16, HULC and Sox4) have been validated. LncRNA SPRY4-IT1 can promote the malignant development of colorectal cancer by targeting epithelial–mesenchymal transition [47]. Chen et al. pointed out that lncRNA CDKN2B-AS1 has been successfully experimentally confirmed [48]. Li et al. found that GAS5 was commonly downregulated in CRC tissues [49]. Zhang et al. pointed out that beta-catenin can bind to KCNQ1OT1 promotor and activate transcription in colon cancer cells [50]. LncRNA BANCR is abnormally expressed in colorectal cancer [51]. LncRNA SNHG16 is regulated by the Wnt pathway in colorectal cancer [52]. LncRNA HULC interacts with miR-613 to regulate colon cancer growth [53]. Lin et al. provided evidence for the clinical significance of overexpressed SOX4 in patients with colon cancer [54]. We list these lncRNAs, their corresponding ranks and corresponding references in Table 4.

Prostate cancer is the most common malignancy among males worldwide [55]. It is the fourth leading cancer in both sexes and the second most common cancer in males. In 2012, about 1.1 million men worldwide with prostate cancer were diagnosed. DMFLDA is applied to predict potential prostate cancer-related lncRNAs. As a result, 7 prostate cancer-related lncRNAs (UHRF1, CCAT1, XIST, TINCR, LINC-ROR, TDRG1, and CRNDE) have been validated by manually mining recent biomedical literature. The results are shown in Table 5. Jazirehi et al. pointed out lncRNA UHRF1 is a master regulator in prostate cancer [56]. LncRNA CCAT1 is a "hot spot" harboring multiple genetic alternations in prostate cancer [57]. Du et al. pointed out that LncRNA XIST acts as a tumor suppressor in prostate cancer through sponging miR-23a to modulate RKIP expression [58]. LncRNA TINCR is associated with clinical progression and serves as tumor suppressive role in prostate cancer [59]. Liu et al. found that curcumin suppresses proliferation and in vitro invasion of human prostate cancer stem cells by lncRNA LINC-ROR [60]. Wang et al. found that lncRNA TDRG1 can regulate the PI3K/Akt/mTOR pathway, which plays an important role in prostate cancer [61]. Ellis et al. pointed out lncRNA CRNDE expression is mildly elevated in prostate cancer [62].

Renal cancer is one of the most common cancers. In 2013, renal cancer was diagnosed in more than 350,000 people worldwide [63]. It causes more than 140,000 deaths per year. Accumulating evidence has shown that lncRNAs play critical roles in the development and progression of renal cancer. DMFLDA is applied to predict potential renal cancer-related lncRNAs. As a result, 7 renal cancer-related lncRNAs (NEAT1, MEG3, GAS5, H19, HOTAIR, AFAP1-AS1, and TDRG1) have been found. LncRNA NEAT1 enhances epithelial-to-mesenchymal transition and chemoresistance in renal cell carcinoma [64]. LncRNA MEG3 induces renal cell carcinoma cells apoptosis [65]. LncRNA GAS5 expression level is significantly lower in renal cell carcinoma samples [66]. Wang et al. found that down-regulated lncRNA H19 inhibits carcinogenesis of renal cell carcinoma [67]. LncRNA HOTAIR activates the Hippo pathway by directly binding to SAV1 in renal cell carcinoma [68]. Lan et al. pointed out that

TABLE 6.
DMFLDA PREDICTED LNCRNAS ASSOCIATED WITH RENAL CANCER (TOP 10) WITH THE CORRESPONDING REFERENCES.

lncRNA	Reference	Rank	Description
NEAT1	Liu et al. (2017)	1	The long non-coding rna neat1 enhances epithelial-to-mesenchymal transition and chemoresistance via the mir-34a/c-met axis in renal cell carcinoma.
MEG3	Wang et al. (2015)	2	Long non-coding rna meg3 induces renal cell carcinoma cells apoptosis by activating the mitochondrial pathway.
GAS5	Seles et al. (2016)	3	Long non-coding RNA GAS5 functions as a tumor suppressor in renal cell carcinoma.
H19	Wang et al. (2015)	4	Down-regulated long non-coding RNA H19 inhibits carcinogenesis of renal cell carcinoma.
HOTAIR	Hu et al. (2017)	5	The long noncoding RNA HOTAIR activates the Hippo pathway by directly binding to SAV1 in renal cell carcinoma.
AFAP1-AS1	Lan et al. (2017)	6	The expression level of AFAP1-AS1 from GSE48352 was higher in papillary renal cell carcinoma (PRCC) than in that of normal controls (P = 0.0318).
TUSC7	Unknown	7	N/A
HMlincRNA717	Unknown	8	N/A
DLEU1	Unknown	9	N/A
TDRG1	Chen et al. (2018)	10	These results suggest TDRG1 regulates cell proliferation, apoptosis and invasion in endometrial cancer–at least in part–by targeting VEGF-A and modulating the expression of proteins regulated by VEGF-A. Studies had already reported that the downregulation of VEGF-A inhibits proliferation, promotes apoptosis and suppresses migration and invasion in renal clear cell carcinoma cells by inhibiting PI3K/Akt expression.

 r more

the expression level of AFAP1-AS1 is higher in renal cell carcinoma than normal controls [69]. Chen et al. found that lncRNA TDRG1 can regulate VEGF-A protein which can inhibit proliferation and promote apoptosis in renal cancer [70].

In summary, 22 cancer-related lncRNAs (8 for colorectal cancer, 7 for prostate cancer and 7 for renal cancer) are checked in the recent biomedical literature. These case studies show that the potential of DMFLDA to infer novel lncRNAs for diseases is confirmed.

4 Conclusions

LncRNAs play important roles in all kinds of fundamental and important biological processes. Identifying disease-related lncRNAs is of great significance in biology for understanding the mechanisms of diseases at the lncRNA level. In this study, we propose a deep matrix factorization-based model DMFLDA to predict potential lncRNA-disease associations. DMFLDA uses deep learning techniques to extract latent vectors of lncRNAs and diseases from their interaction matrix. Then DMFLDA fuses the two vectors into a new vector and use it to perform prediction task. Compared with traditional matrix factorization-based methods, DMFLDA can capture the non-linear, more complex relationships between lncRNAs and diseases. Compared with traditional machine learning-based methods, DMFLDA can obtain more accurate representation learning for lncRNAs and diseases. In order to evaluate the efficiency of our method, we compared DMFLDA with 4 popular computational methods. The LOOCV results demonstrate that DMFLDA performs better than these existing methods. To further evaluate the capacity of DMFLDA, case studies of colorectal cancer, prostate cancer and renal cancer are carried out. 22 cancer-related lncRNAs (8 for colorectal cancer, 7 for prostate cancer, and 7 for renal cancer) are verified by mining recent biomedical literature. These experimental results show that DMFLDA has enough potential to predict novel diseases-related lncRNAs.

In this study, DMFLDA uses lncRNA-disease interactions to construct a deep matrix factorization model. We know that a lot of types of biological information are useful for predicting lncRNA-disease interactions. A possible future work would be to integrate some useful biological information into a deep learning framework to improve the performance. In addition, a better sampling strategy (e.g. using lncRNA or disease similarity to distinguish those instances that are more likely to be negative instances) should be able to improve the performance of the problem; which another direction of our future work.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grants (No. U1909208, No. 61622213 and No. 61728211), the 111 Project (No.B18059), Hunan Provincial Science and Technology Program (2018WK4001), the Fundamental Research Funds for the Central Universities of Central South University (No. 502221903).

References

[1] E. P. Consortium, "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project," *Nature*, vol. 447, no. 7146, pp. 799, 2007.

[2] M. Esteller, "Non-coding RNAs in human disease," *Nature Reviews Genetics*, vol. 12, no. 12, pp. 861, 2011.

[3] Q. Zou, J. Li, Q. Hong, Z. Lin, Y. Wu, H. Shi, and Y. Ju, "Prediction of microRNA-disease associations based on social network analysis methods," *BioMed research international*, vol. 2015, 2015.

[4] Y. Liu, X. Zeng, Z. He, and Q. Zou, "Inferring microRNA-disease associations by random walk on a heterogeneous network with multiple data sources," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 14, no. 4, pp. 905-915, 2016.

[5] W. Tang, S. Wan, Z. Yang, A. E. Teschendorff, and Q. Zou, "Tumor origin detection with tissue-specific miRNA and DNA methylation markers," *Bioinformatics*, vol. 34, no. 3, pp. 398-406, 2017.

[6] X. Zeng, X. Zhang, and Q. Zou, "Integrative approaches for predicting microRNA function and prioritizing disease-related microRNA using biological interaction networks," *Briefings in bioinformatics*, vol. 17, no. 2, pp. 193-203, 2015.

[7] X.-M. Zhao, K.-Q. Liu, G. Zhu, F. He, B. Duval, J.-M. Richer, D.-S. Huang, C.-J. Jiang, J.-K. Hao, and L. Chen, "Identifying cancer-related microRNAs based on gene expression data," *Bioinformatics*, vol. 31, no. 8, pp. 1226-1234, 2014.

[8] N. Hauptman, and D. Glavač, "Long non-coding RNA in cancer," *International journal of molecular sciences*, vol. 14, no. 3, pp. 4655-4669, 2013.

[9] P. Kapranov, J. Cheng, S. Dike, D. A. Nix, R. Duttagupta, A. T. Willingham, P. F. Stadler, J. Hertel, J. Hackermüller, and I. L. Hofacker, "RNA maps reveal new RNA classes and a possible function for pervasive transcription," *Science*, vol. 316, no. 5830, pp. 1484-1488, 2007.

[10] T. R. Mercer, M. E. Dinger, and J. S. Mattick, "Long non-coding RNAs: insights into functions," *Nature reviews genetics*, vol. 10, no. 3, pp. 155, 2009.

[11] M. Guttman, P. Russell, N. T. Ingolia, J. S. Weissman, and E. S. Lander, "Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins," *Cell*, vol. 154, no. 1, pp. 240-251, 2013.

[12] D. Barsyte-Lovejoy, S. K. Lau, P. C. Boutros, F. Khosravi, I. Jurisica, I. L. Andrulis, M. S. Tsao, and L. Z. Penn, "The c-Myc oncogene directly induces the H19 noncoding RNA by allele-specific binding to potentiate tumorigenesis," *Cancer research*, vol. 66, no. 10, pp. 5330-5337, 2006.

[13] S. Lottin, E. Adriaenssens, T. Dupressoir, N. Berteaux, C. Montpellier, J. Coll, T. Dugimont, and J. J. Curgy, "Overexpression of an ectopic H19 gene enhances the tumorigenic properties of breast cancer cells," *Carcinogenesis*, vol. 23, no. 11, pp. 1885-1895, 2002.

[14] C. R. Tessier, G. A. Doyle, B. A. Clark, H. C. Pitot, and J. Ross, "Mammary tumor induction in transgenic mice expressing an RNA-binding protein," *Cancer research*, vol. 64, no. 1, pp. 209-214, 2004.

[15] M. A. Faghihi, F. Modarresi, A. M. Khalil, D. E. Wood, B. G. Information

Sahagan, T. E. Morgan, C. E. Finch, G. S. Laurent III, P. J. Kenny, and C. Wahlestedt, "Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of β-secretase," *Nature medicine*, vol. 14, no. 7, pp. 723, 2008.

[16] X. Chen, and G.-Y. Yan, "Novel human lncRNA–disease association inference based on lncRNA expression profiles," *Bioinformatics*, vol. 29, no. 20, pp. 2617-2624, 2013.

[17] T. Zhao, J. Xu, L. Liu, J. Bai, C. Xu, Y. Xiao, X. Li, and L. Zhang, "Identification of cancer-related lncRNAs through integrating genome, regulome and transcriptome features," *Molecular BioSystems*, vol. 11, no. 1, pp. 126-136, 2015.

[18] W. Lan, M. Li, K. Zhao, J. Liu, F.-X. Wu, Y. Pan, and J. Wang, "LDAP: a web server for lncRNA-disease association prediction," *Bioinformatics*, vol. 33, no. 3, pp. 458-460, 2016.

[19] G. Fu, J. Wang, C. Domeniconi, and G. Yu, "Matrix factorization-based data fusion for the prediction of lncRNA–disease associations," *Bioinformatics*, vol. 34, no. 9, pp. 1529-1537, 2017.

[20] C. Lu, M. Yang, F. Luo, F.-X. Wu, M. Li, Y. Pan, Y. Li, and J. Wang, "Prediction of lncRNA-disease associations based on inductive matrix completion," *Bioinformatics*, vol. 1, pp. 8, 2018.

[21] J. Sun, H. Shi, Z. Wang, C. Zhang, L. Liu, L. Wang, W. He, D. Hao, S. Liu, and M. Zhou, "Inferring novel lncRNA–disease associations based on a random walk model of a lncRNA functional similarity network," *Molecular BioSystems*, vol. 10, no. 8, pp. 2074-2081, 2014.

[22] Q. Yao, L. Wu, J. Li, L. guang Yang, Y. Sun, Z. Li, S. He, F. Feng, H. Li, and Y. Li, "Global prioritizing disease candidate lncRNAs via a multi-level composite network," *Scientific reports*, vol. 7, pp. 39516, 2017.

[23] X. Chen, "KATZLDA: KATZ measure for the lncRNA-disease association prediction," *Scientific reports*, vol. 5, pp. 16840, 2015.

[24] M. Zhou, X. Wang, J. Li, D. Hao, Z. Wang, H. Shi, L. Han, H. Zhou, and J. Sun, "Prioritizing candidate disease-related long non-coding RNAs by walking on the heterogeneous lncRNA and disease network," *Molecular bioSystems*, vol. 11, no. 3, pp. 760-769, 2015.

[25] J. Zhang, Z. Zhang, Z. Chen, and L. Deng, "Integrating multiple heterogeneous networks for novel lncRNA-disease association inference," *IEEE/ACM transactions on computational biology and bioinformatics*, 2017.

[26] M.-X. Liu, X. Chen, G. Chen, Q.-H. Cui, and G.-Y. Yan, "A computational framework to infer human disease-associated long noncoding RNAs," *PloS one*, vol. 9, no. 1, pp. e84408, 2014.

[27] X. Chen, "Predicting lncRNA-disease associations and constructing lncRNA functional similarity network based on the information of miRNA," *Scientific reports*, vol. 5, pp. 13186, 2015.

[28] M. Zeng, M. Li, Z. Fei, Y. Yu, Y. Pan, and J. Wang, "Automatic ICD-9 coding via deep transfer learning," *Neurocomputing*, vol. 324, pp. 43-50, 2019.

[29] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks." pp. 1097-1105.

[30] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality." pp. 3111-3119.

[31] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[32] M. Li, Z. Fei, M. Zeng, F. Wu, Y. Li, Y. Pan, and J. Wang, "Automated ICD-9 Coding via A Deep Learning Approach," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, no. 1, pp. 1-1, 2018.

[33] M. Zeng, M. Li, Z. Fei, F. Wu, Y. Li, Y. Pan, and J. Wang, "A deep learning framework for identifying essential proteins by integrating multiple types of biological information," *IEEE/ACM transactions on computational biology and bioinformatics*, 2019.

[34] M. Zeng, F. Zhang, F.-X. Wu, Y. Li, J. Wang, and M. Li, "Protein–protein interaction site prediction through combining local and global features with deep neural networks," *Bioinformatics*, 2019.

[35] F. Zhang, H. Song, M. Zeng, Y. Li, L. Kurgan, and M. Li, "DeepFunc: A Deep Learning Framework for Accurate Prediction of Protein Functions from Protein Sequences and Interactions," *Proteomics*, pp. 1900019, 2019.

[36] H.-J. Xue, X. Dai, J. Zhang, S. Huang, and J. Chen, "Deep Matrix Factorization Models for Recommender Systems." pp. 3203-3209.

[37] M. van Baalen, "Deep Matrix Factorization for Recommendation," 2016.

[38] L. Ding, M. Wang, D. Sun, and A. Li, "TPGLDA: Novel prediction of associations between lncRNAs and diseases via lncRNA-disease-gene tripartite graph," *Scientific reports*, vol. 8, no. 1, pp. 1065, 2018.

[39] M. Zeng, B. Zou, F. Wei, X. Liu, and L. Wang, "Effective prediction of three common diseases by combining SMOTE with Tomek links technique for imbalanced medical data." pp. 225-228.

[40] M. Zeng, M. Li, Z. Fei, F.-X. Wu, Y. Li, and Y. Pan, "A deep learning framework for identifying essential proteins based on protein-protein interaction network and gene expression data." pp. 583-588.

[41] M. Zeng, M. Li, F.-X. Wu, Y. Li, and Y. Pan, "DeepEP: a deep learning framework for identifying essential proteins," *BMC Bioinformatics*, vol. 20, no. 16, pp. 506, 2019.

[42] G. Chen, Z. Wang, D. Wang, C. Qiu, M. Liu, X. Chen, Q. Zhang, G. Yan, and Q. Cui, "LncRNADisease: a database for long-non-coding RNA-associated diseases," *Nucleic acids research*, vol. 41, no. D1, pp. D983-D986, 2012.

[43] Z. Lu, K. BRETONNEL COHEN, and L. Hunter, "GeneRIF quality assurance as summary revision," *Biocomputing 2007*, pp. 269-280: World Scientific, 2007.

[44] S. Ning, J. Zhang, P. Wang, H. Zhi, J. Wang, Y. Liu, Y. Gao, M. Guo, M. Yue, and L. Wang, "Lnc2Cancer: a manually curated database of experimentally supported lncRNAs associated with various human cancers," *Nucleic Acids Research*, vol. 44, no. Database issue, pp. D980-D985, 2016.

[45] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, and M. Isard, "Tensorflow: a system for large-scale machine learning." pp. 265-283.

[46] P. Favoriti, G. Carbone, M. Greco, F. Pirozzi, R. E. M. Pirozzi, "Designing heterogeneous ensembles with contextual information"

and F. Corcione, "Worldwide burden of colorectal cancer: a review," *Updates in surgery,* vol. 68, no. 1, pp. 7-11, 2016.

[47] D. Cao, Q. Ding, W. Yu, M. Gao, and Y. Wang, "long noncoding rna SPRY4-IT1 promotes malignant development of colorectal cancer by targeting epithelial–mesenchymal transition," *OncoTargets and therapy,* vol. 9, pp. 5417, 2016.

[48] X. Chen, C. C. Yan, X. Zhang, and Z.-H. You, "Long non-coding RNAs and complex diseases: from experimental results to computational models," *Briefings in bioinformatics,* vol. 18, no. 4, pp. 558-576, 2016.

[49] Y. Li, Y. Li, S. Huang, K. He, M. Zhao, H. Lin, D. Li, J. Qian, C. Zhou, and Y. Chen, "Long non-coding RNA growth arrest specific transcript 5 acts as a tumour suppressor in colorectal cancer by inhibiting interleukin-10 and vascular endothelial growth factor expression," *Oncotarget,* vol. 8, no. 8, pp. 13690, 2017.

[50] C. Zhang, S. Du, and L. Cao, "Long non-coding RNA KCNQ1OT1 promotes osteosarcoma progression by increasing β-catenin activity," *RSC Advances,* vol. 8, no. 66, pp. 37581-37589, 2018.

[51] S. Su, J. Gao, T. Wang, J. Wang, H. Li, and Z. Wang, "Long non-coding RNA BANCR regulates growth and metastasis and is associated with poor prognosis in retinoblastoma," *Tumor Biology,* vol. 36, no. 9, pp. 7205-7211, 2015.

[52] L. L. Christensen, K. True, M. P. Hamilton, M. M. Nielsen, N. D. Damas, C. K. Damgaard, H. Ongen, E. Dermitzakis, J. B. Bramsen, and J. S. Pedersen, "SNHG16 is regulated by the Wnt pathway in colorectal cancer and affects genes involved in lipid metabolism," *Molecular oncology,* vol. 10, no. 8, pp. 1266-1282, 2016.

[53] Y. Dong, M.-H. Wei, J.-G. Lu, and C.-Y. Bi, "Long non-coding RNA HULC interacts with miR-613 to regulate colon cancer growth and metastasis through targeting RTKN," *Biomedicine & Pharmacotherapy,* vol. 109, pp. 2035-2042, 2019.

[54] C.-M. Lin, C.-L. Fang, Y.-C. Hseu, C.-L. Chen, J.-W. Wang, S.-L. Hsu, M.-D. Tu, S.-T. Hung, C. Tai, and Y.-H. Uen, "Clinical and prognostic implications of transcription factor SOX4 in patients with colon cancer," *PLoS One,* vol. 8, no. 6, pp. e67128, 2013.

[55] M. N. Bashir, "Epidemiology of prostate cancer," *Asian Pac J Cancer Prev,* vol. 16, no. 13, pp. 5137-41, 2015.

[56] A. R. Jazirehi, D. Arle, and P. B. Wenn, "UHRF1: a master regulator in prostate cancer," *Epigenomics,* vol. 4, no. 3, pp. 251, 2012.

[57] I. Mizrahi, H. Mazeh, R. Grinbaum, N. Beglaibter, M. Wilschanski, V. Pavlov, M. Adileh, A. Stojadinovic, I. Avital, and A. O. Gure, "Colon cancer associated transcript-1 (CCAT1) expression in adenocarcinoma of the stomach," *Journal of Cancer,* vol. 6, no. 2, pp. 105, 2015.

[58] Y. Du, X.-D. Weng, L. Wang, X.-H. Liu, H.-C. Zhu, J. Guo, J.-Z. Ning, and C.-C. Xiao, "LncRNA XIST acts as a tumor suppressor in prostate cancer through sponging miR-23a to modulate RKIP expression," *Oncotarget,* vol. 8, no. 55, pp. 94358, 2017.

[59] L. Dong, H. Ding, Y. Li, D. Xue, and Y. Liu, "LncRNA TINCR is associated with clinical progression and serves as tumor suppressive role in prostate cancer," *Cancer management and research,* vol. 10, pp. 2799, 2018.

[60] T. Liu, H. Chi, J. Chen, C. Chen, Y. Huang, H. Xi, J. Xue, and Y. Si, "Curcumin suppresses proliferation and in vitro invasion of human prostate cancer stem cells by ceRNA effect of miR-145 and lncRNA-ROR," *Gene,* vol. 631, pp. 29-38, 2017.

[61] Y. Wang, Y. Gan, Z. Tan, J. Zhou, R. Kitazawa, X. Jiang, Y. Tang, and J. Yang, "TDRG1 functions in testicular seminoma are dependent on the PI3K/Akt/mTOR signaling pathway," *OncoTargets and therapy,* vol. 9, pp. 409, 2016.

[62] B. C. Ellis, P. L. Molloy, and L. D. Graham, "CRNDE: a long non-coding RNA involved in cancer, neurobiology, and development," *Frontiers in genetics,* vol. 3, pp. 270, 2012.

[63] S. Zhou, J. Wang, and Z. Zhang, "An emerging understanding of long noncoding RNAs in kidney cancer," *Journal of cancer research and clinical oncology,* vol. 140, no. 12, pp. 1989-1995, 2014.

[64] F. Liu, N. Chen, Y. Gong, R. Xiao, W. Wang, and Z. Pan, "The long non-coding RNA NEAT1 enhances epithelial-to-mesenchymal transition and chemoresistance via the miR-34a/c-Met axis in renal cell carcinoma," *Oncotarget,* vol. 8, no. 38, pp. 62927, 2017.

[65] M. Wang, T. Huang, G. Luo, C. Huang, X.-y. Xiao, L. Wang, G.-s. Jiang, and F.-q. Zeng, "Long non-coding RNA MEG3 induces renal cell carcinoma cells apoptosis by activating the mitochondrial pathway," *Journal of Huazhong University of Science and Technology [Medical Sciences],* vol. 35, no. 4, pp. 541-545, 2015.

[66] M. Seles, G. Hutterer, T. Kiesslich, K. Pummer, I. Berindan-Neagoe, S. Perakis, D. Schwarzenbacher, M. Stotz, A. Gerger, and M. Pichler, "Current insights into long non-coding RNAs in renal cell carcinoma," *International journal of molecular sciences,* vol. 17, no. 4, pp. 573, 2016.

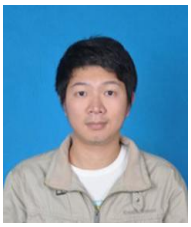
[67] L. Wang, Y. Cai, X. Zhao, X. Jia, J. Zhang, J. Liu, H. Zhen, T. Wang, X. Tang, and Y. Liu, "Down-regulated long non-coding RNA H19 inhibits carcinogenesis of renal cell carcinoma," *Neoplasma,* vol. 62, no. 3, pp. 412-418, 2015.

[68] G. Hu, B. Dong, J. Zhang, W. Zhai, T. Xie, B. Huang, C. Huang, X. Yao, J. Zheng, and J. Che, "The long noncoding RNA HOTAIR activates the Hippo pathway by directly binding to SAV1 in renal cell carcinoma," *Oncotarget,* vol. 8, no. 35, pp. 58654, 2017.

[69] H. Lan, J. Zeng, G. Chen, and H. Huang, "Survival prediction of kidney renal papillary cell carcinoma by comprehensive LncRNA characterization," *Oncotarget,* vol. 8, no. 67, pp. 110811, 2017.

[70] S. Chen, L.-l. Wang, K.-x. Sun, Y. Liu, X. Guan, Z.-h. Zong, and Y. Zhao, "LncRNA TDRG1 enhances tumorigenicity in endometrial carcinoma by binding and targeting VEGF-A protein," *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease,* 2018.

Min Zeng received the B.S. degree from Lanzhou University in 2013, and the M.S. degree from Central South University in 2016. He is currently working toward the PhD degree in the School of Computer Science and Engineering, Central South University, China. His research interests include bioinformatics, machine learning and deep learning.



Chengqian Lu received the B.S. degree from Xiangtan University in 2009, and the M.S. degree from Yunnan University in 2011. He is a PhD candidate in the School of Computer Science and Engineering, Central South University, China. His current research interests include bioinformatics, machine learning and deep learning.

Zhihui Fei received his BSc degrees in Hubei Normal University, China in 2015. He is currently a postgraduate student in Bioinformatics at Central South University. His currently research interests include bioinformatics, medical data mining and deep learning.

Fang-Xiang Wu (M'06-SM'11) received the B.Sc. degree and the M.Sc. degree in applied mathematics, both from Dalian University of Technology, Dalian, China, in 1990 and 1993, respectively, the first Ph.D. degree in control theory and its applications from Northwestern Polytechnical University, Xi'an, China, in 1998, and the second Ph.D. degree in biomedical engineering from University of Saskatchewan (U of S), Saskatoon, Canada, in 2004. During 2004-2005, he worked as a Postdoctoral Fellow in the Laval University Medical Research Center (CHUL), Quebec City, Canada. He is currently a Professor of the Division of Biomedical Engineering and the Department of Mechanical Engineering at the U of S. His current research interests include computational and systems biology, genomic and proteomic data analysis, biological system identification and parameter estimation, applications of control theory to biological systems. Dr. Wu is serving as the editorial board member of five international journals, the guest editor of several international journals, and as the program committee chair or member of several international conferences. He has also reviewed papers for many international journals.

Yaohang Li received the M.S. and Ph.D. degrees in computer science from Florida State University, Tallahassee, FL, USA, in 2000 and 2003, respectively. He is an Associate Professor in the Department of Computer Science at Old Dominion University, Norfolk, VA, USA. Hisresearch interests are in computational biology, Monte Carlo methods, and scientific computing. After graduation, he worked at Oak Ridge National Laboratory as a Research Associate for a short period. Before joining ODU, he was an Associate Professor in the Computer Science Department at North Carolina A&T State University.

Jianxin Wang received the BEng and MEng degrees in computer engineering from Central South University, China, in 1992 and 1996, respectively, and the PhD degree in computer science from Central South University, China, in 2001. He is the dean and a professor in School of Computer Science and Engineering, Central South University, Changsha, Hunan, P.R. China. His current research interests include algorithm analysis and optimization, parameterized algorithm, bioinformatics and computer network. He has published more than 150 papers in various International journals and refereed conferences. He is a senior member of the IEEE.

Min Li received the PhD degree in Computer Science from Central South University, China, in 2008. She is currently a Professor at the School of Computer Science and Engineering, Central South University, Changsha, Hunan, P.R. China. Her research interests include computational biology, systems biology and bioinformatics. She has published more than 80 technical papers in refereed journals such as Bioinformatics, IEEE/ACM Transactions on Computational Biology and Bioinformatics, Proteomics, and conference proceedings such as BIBM, GIW and ISBRA. According to Google scholar, her paper citations are more than 4000 and H-index is 34.