

RESEARCH

Open Access



SVDNLDA: predicting lncRNA-disease associations by Singular Value Decomposition and *node2vec*

Jianwei Li^{1,2}, Jianing Li^{1,2}, Mengfan Kong^{1,2}, Duanyang Wang^{1,2}, Kun Fu^{1,2*} and Jiangcheng Shi^{3*}

*Correspondence:

fukun@hebut.edu.cn;
shijiangcheng@tiangong.edu.cn

¹ Institute of Computational Medicine, School of Artificial Intelligence, Hebei University of Technology, Tianjin 300401, China

³ School of Life Sciences, Tiangong University, Tianjin 300387, China
Full list of author information is available at the end of the article

Abstract

Background: Numerous studies on discovering the roles of long non-coding RNAs (lncRNAs) in the occurrence, development and prognosis progresses of various human diseases have drawn substantial attentions. Since only a tiny portion of lncRNA-disease associations have been properly annotated, an increasing number of computational methods have been proposed for predicting potential lncRNA-disease associations. However, traditional predicting models lack the ability to precisely extract features of biomolecules, it is urgent to find a model which can identify potential lncRNA-disease associations with both efficiency and accuracy.

Results: In this study, we proposed a novel model, SVDNLDA, which gained the linear and non-linear features of lncRNAs and diseases with Singular Value Decomposition (SVD) and *node2vec* methods respectively. The integrated features were constructed from connecting the linear and non-linear features of each entity, which could effectively enhance the semantics contained in ultimate representations. And an XGBoost classifier was employed for identifying potential lncRNA-disease associations eventually.

Conclusions: We propose a novel model to predict lncRNA-disease associations. This model is expected to identify potential relationships between lncRNAs and diseases and further explore the disease mechanisms at the lncRNA molecular level.

Keywords: lncRNA-disease association prediction, Singular Value Decomposition, *node2vec*, Network representation learning, XGBoost classifier

Background

Since the central dogma of molecular biology was proposed, RNA has been treated as an intermediary between protein-coding gene and protein. However, protein-coding genes account for only ~ 1.5% of the human genome, and more than 98% of the human genome cannot encode proteins [1–3]. Most non-coding genes would be transcribed into non-coding RNAs (ncRNAs). As their names imply, ncRNAs cannot be directly translated into proteins, so they were often considered as the "noise" of genome transcription without any biological functions for decades. According to the lengths of nucleotide



© The Author(s). 2021. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

sequences, ncRNAs can be further divided into small ncRNAs (< 200 nucleotides) and long ncRNAs (> 200 nucleotides) [4, 5]. Following the discovery of lncRNA H19 and XIST in the early 1990s [6, 7], associated with the rapid developments of scientific methodologies and experimental techniques, researchers have identified thousands of lncRNAs in eukaryotes ranging from nematodes to humans [8, 9]. Abundant evidences have demonstrated that lncRNAs play important roles in many fundamental and critical biological processes, such as transcriptional and post-transcriptional regulation, epigenetic regulation and chromosome dynamics [10–14]. Previous studies showed that the mutation or dysregulation of lncRNAs are closely related with a variety of human diseases. For instance, MALAT1, also known as NEAT2, was found upregulated in non-small cell lung cancer tissues and could be served as an early prognostic biomarker [15]; lncRNA HOTAIR had been explored as a potential biomarker on the detection of hepatocellular carcinoma relapse [16].

The complex and precise regulatory functions of lncRNAs have largely explained the complexity of genome and opened a new chapter for scientists to deeply understand the diversity of living organisms from the perspective on gene expression regulatory network. However, the exact mechanisms behind these various regulative relationships remain to be further explored; the general characteristics of lncRNAs, such as the relationships between their spatial structures and functions, the realization of transcriptional regulation, and the molecular level mechanisms in various biological processes or diseases, are still unknown. The identification of lncRNA-disease associations can not only help us better understand the underlying mechanisms of lncRNAs in various human diseases, but also accelerate the discovery of potential biomarkers which may benefit the diagnosis, treatment, prognosis of many complex diseases. The exploration on the association between lncRNA and disease has attracted more and more researchers' attention nowadays, which has become a prevalent topic in the current research field of lncRNA. Due to the number of newly discovered lncRNAs is growing rapidly every year, identifying lncRNA-disease association purely based on clinical information and biological experiments has encountered bottlenecks for their enormous consume of time and cost, and their disability to predict the associations of unrecorded diseases or lncRNAs, which undoubtedly limits the development of the lncRNA related studies. However, computational methods based on biological data can rapidly and efficiently quantify the correlation probability of interested lncRNA-disease pairs automatically, which can significantly reduce the time and cost of biological experiments. Therefore, it is a significant and urgent task to develop efficient and robust computational methods that are capable for predicting potential lncRNA-disease associations and providing candidates for future experimental verification.

Many researchers have proposed numerous algorithms and models for predicting potential lncRNA-disease association relationships over the years. All these methods could be broadly divided into three groups: biological network-based methods, machine learning-based methods and others. Based on the hypothesis that lncRNAs with similar functions may be more likely to be associated with diseases with similar phenotypes [17], a significant number of different biological network-based methods have been proposed by integrating multi-source biological information networks to detect potential disease-related lncRNAs. Sun et al. [18] proposed a global network-based computing

method, RWRLNCD. By integrating a lncRNA-disease association network and a disease similarity network into a lncRNA functional similarity network, RWRLNCD adopted the Random Walk with Restart (RWR) algorithm on the constructed lncRNA functionally similar network to conduct predictions. Yao et al. [19] proposed a predictive model named LNCPrNet, which was based on a multi-layer composite network fusing different data of phenotypic-phenotypic interactions, lncRNA-lncRNA interactions and gene-gene interactions with disease-ncRNA relationships. The RWR algorithm was applied to predict potential lncRNA-disease associations. LNCPrNet could still achieve a decent performance when the known lncRNA-disease association data was insufficient, which may largely thank to the fact that the multi-layer composite network interacted with abundant information offsetted the insufficient with one particular type of data. Ding et al. [20] came up with a model named TPGLDA in which built a lncRNA-disease-gene tripartite graph and applied a resource allocation algorithm to obtain the promising lncRNA-disease associations. Zhao et al. [21] built a multi-heterogeneous network which integrated the lncRNA functional similarity network, genetic similarity network, disease semantic similarity network and association networks among these three kinds of biological entries, subsequently realized the prediction of underlying lncRNA-disease associations through the RWR algorithm on their heterogeneous network. Xie et al. [22] adopted unbalance bi-random walk in their heterogeneous network to reconstruct the lncRNA-disease association matrix, which reflected the latent lncRNA-disease associations. After that, they proposed a NCPHLDA model [23], which constructed two cosine similarity networks for all lncRNAs and diseases separately, and combined the network consistency projection score for each similarity network as the associated probability of corresponding lncRNA-disease pairs. Most of these biological network-based methods adopted random walk-based algorithms on the established heterogeneous networks, which essentially takes the underlying topology information of nodes in the heterogeneous networks as the basis for the potential association prediction. The predicted effects of network-based methods heavily depend on whether the built network could accurately and comprehensively reflect the interactions among real biomolecules. Meanwhile, the rigid neighborhood relationship utilized by the random walk algorithm or its derivations limits the information richness of molecular features.

In recent years, machine learning and deep learning techniques have been widely adopted in lncRNA-disease association predictions. Most of machine learning methods for disease-related lncRNA candidate selection typically train classifiers with the acquired features of experimentally confirmed lncRNA-disease associations and interested candidates, then rank the candidating associations according to the classification results. Chen et al. [17] came up with a calculating model, LRLSLDA (Laplacian Regularized Least Squares for lncRNA-Disease Association), based on the “guilt by association” assumption that similar diseases tend to be associated with lncRNAs which possess similar functions. They developed a semi-supervised learning framework to predict potential disease-lncRNA associations. However, there are too many parameters involved in their model, and how to adjust parameters was not well addressed. In addition, the same lncRNA-disease pairs may get different scores from the lncRNA space and the disease space respectively, how to properly combining these scores is a tricky problem. Liu et al. [24] designed a computational model by integrating known human disease

genes, human lncRNAs and gene expression profiles without relying on any known human lncRNA-disease relationships. However, this model could not predict disease-associated lncRNAs which have no associated gene records. Guo et al. [25] integrated the Gaussian interaction profile kernel similarity of lncRNAs and diseases with disease semantic similarity, and utilized an autoencoder getting lower-dimensional features of lncRNA-disease pairs. Finally, a rotating forest classifier was adopted to gain the prediction results. Beyond that, several deep learning-based models have been developed in lncRNA-disease prediction field. Zeng et al. [26] initially combined matrix factorization method with a two-hidden-layer neural network architecture to capture the linear and non-linear features of lncRNAs and diseases respectively. Subsequently, they proposed a deep learning framework named DMFLDA [27], which adopted deep matrix factorization to learn the represents of lncRNAs and diseases. Besides, they also proposed a SDLDA model [28] mixed matrix factorization method with neural network framework to extract different features of lncRNAs and diseases.

In addition to biological networks and machine learning methods, plenty of statistical methods are also adopted to predict latent lncRNA-disease associations. Chen et al. [29] proposed a HGLDA model based on hypergeometric distribution, where the functional similarity of lncRNA was calculated by integrating disease semantic similarity, miRNA-disease association, and miRNA-lncRNA interaction. By testing whether the number of the common miRNAs shared by the disease and the lncRNA which were in the same lncRNA-disease pair exceeded beyond some threshold, HGLDA performed hypergeometric distribution tests for each lncRNA-disease pair. Lu et al. [30] proposed a matrix factorization-based model, SIMCLDA. According to known lncRNA-disease, gene-disease, gene-gene interactions and the functional similarities of diseases, the Gaussian interaction kernel of lncRNAs was calculated, the matrix decomposition method was introduced to predict the potential lncRNA-disease associations. However, it did not tackle the problem of data sparsity and further studies are needed to improve its performance. Apart from statistical methods, there are still a lot of novel algorithms could be applied for potential association predictions. For example, Fan et al. [31] introduced graph convolutional matrix completion to implement potential lncRNA-disease associations. Fusing verified lncRNA-disease associations and similarity data, they constructed an encoder-decoder model to learn nodes embeddings and score associations respectively.

In this paper, we propose an integrated feature extraction model, Singular Value Decomposition SVD and Node2Vec based lncRNA-Disease Association prediction model (SVDNVLDA), to predict potential lncRNA-disease associations. The rest of this paper is arranged as follows:

The results and discussions section exhibits the influences of hyperparameters in SVDNVLDA, the results of model comparison, robustness test and case studies, as well as an in-depth analysis of the limitations of SVDNVLDA and further improvement directions.

The conclusion section overviews the workflow of SVDNVLDA, and its first-class prediction capability for practical applications.

The methods section introduces the acquisition and preprocessing of experimental data, the prediction process of SVDNVLDA, and the theoretical details of SVD and node2vec methods involved in our model.

Table 1 The ACC results of different features on classifiers

	SVD	N2V16	SN2V16	N2V32	SN2V32	N2V64	SN2V64	N2V128	SN2V128
LR	0.9207	0.9381	0.9404	0.9357	0.9370	0.9348	0.9366	0.9389	0.9374
NB	0.8327	0.9189	0.8456	0.9079	0.8559	0.9046	0.8644	0.8978	0.8795
RF	0.9261	0.9288	0.9308	0.9248	0.9303	0.9246	0.9304	0.9257	0.9385
ADB	0.9307	0.9361	0.9376	0.9331	0.9375	0.9289	0.9378	0.9281	0.9350
XGB	0.9383	0.9400	0.9460	0.9392	0.9454	0.9365	0.9452	0.9364	0.9444

Table 2 The MCC results of different features on classifiers

	SVD	N2V16	SN2V16	N2V32	SN2V32	N2V64	SN2V64	N2V128	SN2V128
LR	0.8412	0.8766	0.8812	0.8717	0.8743	0.8698	0.8736	0.8779	0.8750
NB	0.6719	0.8391	0.6956	0.8717	0.7150	0.8116	0.7317	0.7994	0.7613
RF	0.8527	0.8579	0.8619	0.8500	0.8607	0.8496	0.8609	0.8517	0.8618
ADB	0.8617	0.8705	0.8754	0.8667	0.8752	0.8580	0.8758	0.8566	0.8700
XGB	0.8770	0.8803	0.8922	0.8789	0.8920	0.8730	0.8906	0.8730	0.8891

Results and discussions

Evaluation metrics

Except for special instructions, all the numerical experimental results involved in this paper were generated under tenfold cross-validations. The evaluation metrics used in classifier selection and parameter adjustment processes contained Accuracy (*Acc*), Sensitivity (*Sen*), Specificity (*Spec*), Precision (*Prec*), and Matthews correlation coefficient (*MCC*) [32, 33]. In contrast experiments, the average AUC values and the AUPR values of ten testing sets of each model were gained and the corresponding ROC curves and PR curves were drawn through the results of tenfold cross-validations [34, 35].

Classifier selection and parameter tuning

After gaining the linear feature matrixes U and V^T based on SVD, we found a huge decay gap from 10^{-1} to 10^{-14} between the 173rd and the 174th dimensions of the importance matrix Σ (Additional file 1). In the light of principle of SVD, the linear features of entities were mainly focused on the top 173 dimensions. Therefore, the linear feature vectors of lncRNA and disease were fixed to 173 dimensions. As node2vec is a highly encapsulated node representation learning method, most of the inner parameters were kept constant and the hyperparameters acted as the dimensions of nonlinear vectors in our model. The 16-, 32-, 64-, and 128-dimensional nonlinear feature representations were obtained, respectively.

In the selection process of machine learning classifiers, Linear Regression (LR), Naive Bayes (NB) [36], Random Forest (RF) [37], AdaBoost (ADB) [38] and XGB (XGBoost) [39] were tested based on different integrated features, respectively. The results of ACC and MCC values of all classifiers are shown in Tables 1 and 2. The column named “SVD” represents the features extracted based on single SVD method.

Analogously, “N2V16” represents the 16-dimensional features extracted based on node2vec, “SN2V16” represents the integrated features combined with SVD features and 16 dimensional node2vec, and so on. For results on other evaluation indexes Sen, Spec and Prec, refer to Additional file 2, Additional file 3 and Additional file 4 respectively. All above classifiers were imported from scikit-learn library and implemented on Python, all inner-classifier parameters were set as defaults.

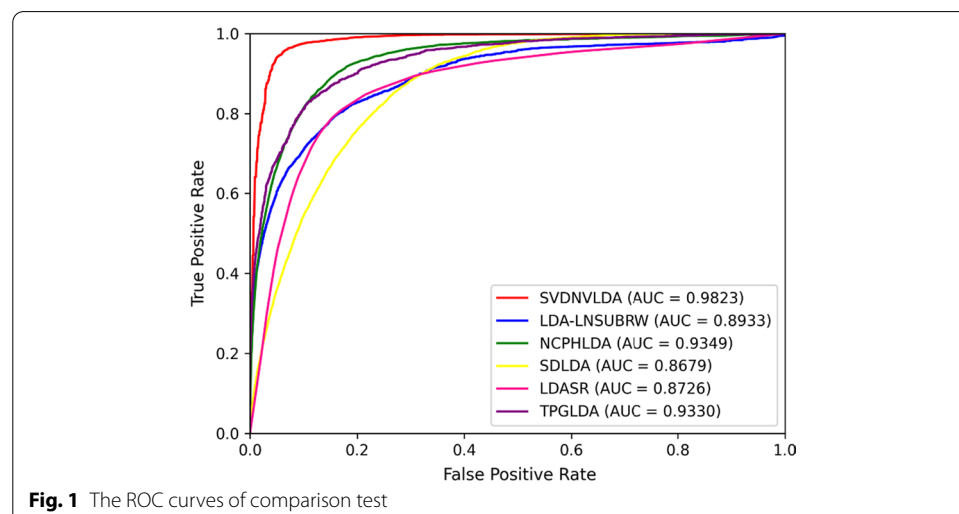
As known from Tables 1 and 2, the combination of linear features and 16-dimensional node2vec features obtained the optimal classification results in the XGBoost classifier (bolded in Tables 1, 2). Moreover, in most classifiers, prediction results based on integrated features were better than single linear feature prediction results and corresponding nonlinear feature prediction results, which demonstrated that the combination of SVD and node2vec does enhance the expression of integrated feature vectors in majority of classifiers.

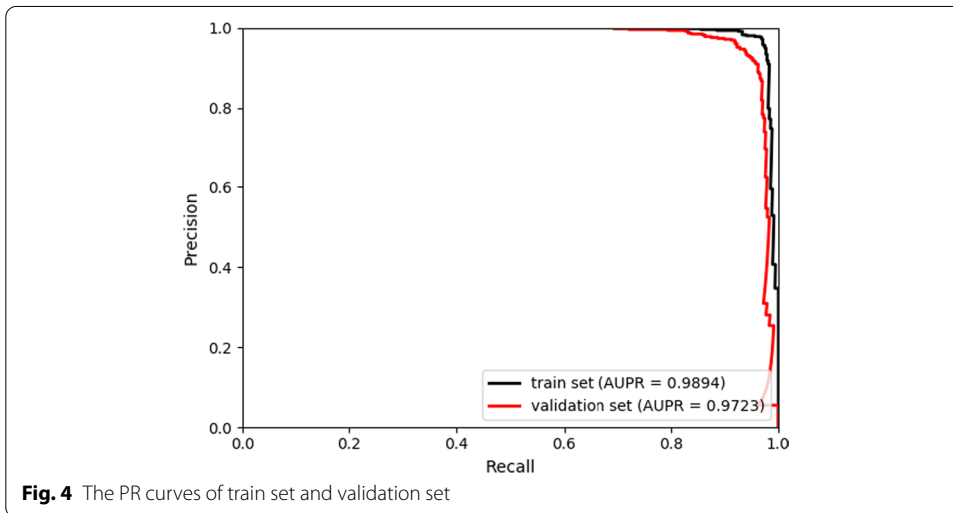
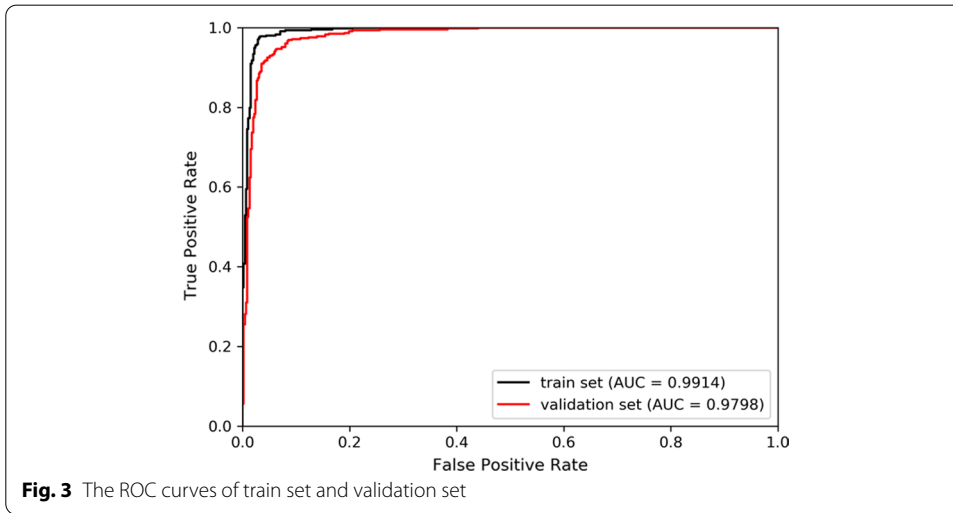
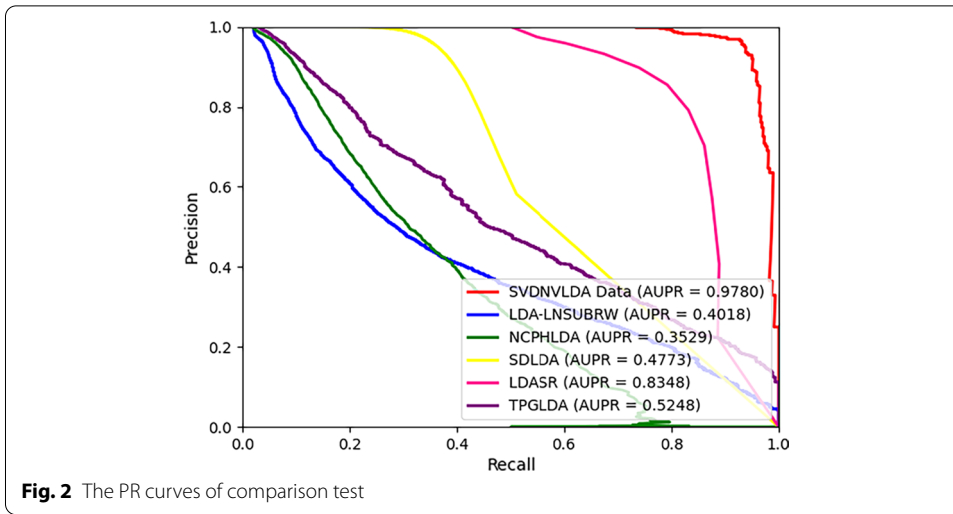
Model contrast

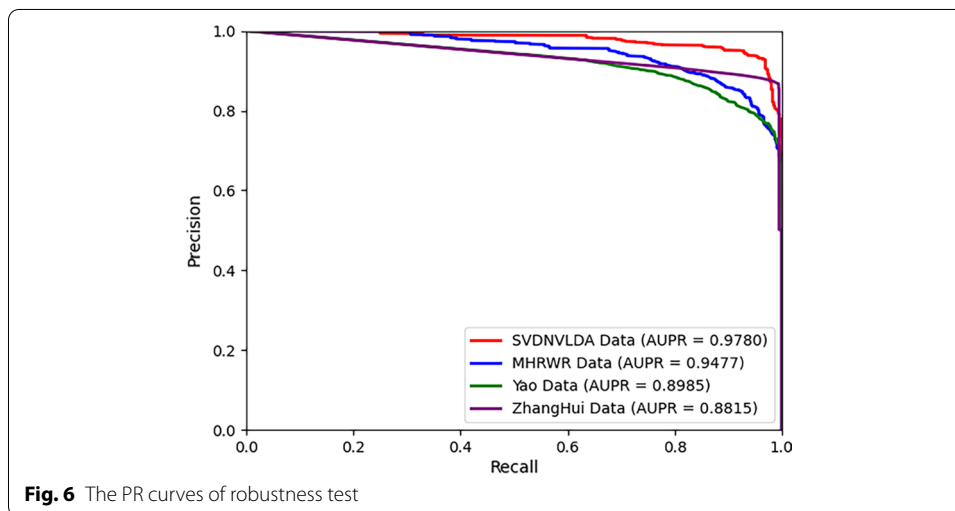
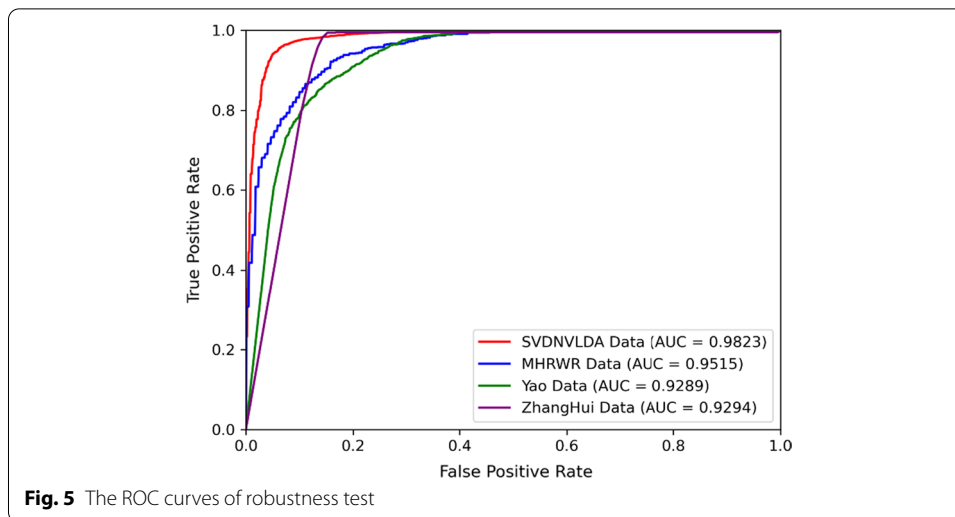
After the model construction, we compared the proposed model with five state-of-the-art lncRNA-disease prediction methods: LDASR [25], LDA-LNSUBRRW [22], NCPHLDA [23], SDLDA [28], and TPGLDA [20]. The ROC and PR curves under ten-fold cross-validations as well as relevant AUC and AUPR values are shown in Figs. 1 and 2 respectively.

Just as shown in Figs. 1 and 2, both the AUC value and AUPR value of SVDNVLDA are the highest among tested models, which indicated that the outperformance of SVDNVLDA. In terms of AUC, compared with NCPHLDA model, which gained the best result in contrast group, our model also improved the AUC value by about 5%. Moreover, the excellent AUPR value manifested that our model also has first-class classification ability on unbalance data sets.

Since all parameters of XGBoost classifier were set as defaults, to testify whether the AUC and AUPR results of SVDNVLDA is overfitted, we further separated 10% samples as validation set and trained classifier without leveraging the validation set. The ROC







and PR curves of the train set and the validation set were exhibited in Figs. 3 and 4 respectively. SVDNLDA achieved remarkable results with AUC of 0.9798 and AUPR of 0.9723 on the validation set, and it was not a result of overfitting.

Robustness testing

The robustness of the predictive model is that the predictive model can give a stable performance for data sets on different scales. For evaluating the robustness of SVDNLDA, we applied it on three varying scale data sets, which had been adopted by other open-source lncRNA-disease association identification models. Similarly, under ten-fold cross validations, the ROC and the PR curves of SVDNLDA on these data sets are plotted in Figs. 5 and 6 respectively. The data set used in Yao's model [19] includes 2697 lncRNA-disease associations, 1002 lncRNA-miRNA associations, and 13,562 miRNA-disease associations. And the data leveraged in Zhanghui's model [40] contains 1151 lncRNA-disease associations, 10,102 lncRNA-miRNA associations and 4634

miRNA-disease associations. While, it is worth mentioning that miRNA entities were replaced with genes in the data set of MHRWR [21], which included 264 lncRNA-gene associations, 855 lncRNA-disease associations, and 9997 gene-disease associations. The experimental test results yielded that SVDNLDA achieved excellent prediction results on all data sets, in particular, the prediction results after replacing miRNAs with other biological entities were still fine in the MHRWR model data. All these results suggested that SVDNLDA can be flexible to accommodate data in different scales or even different contents.

Case studies

To further evaluate the performance of SVDNLDA model in practical applications, we selected lung cancer, breast cancer and pancreatic cancer as case studies. The general processes of each of case studies were as following: first, all lncRNA-disease association data and the same number of negative samples were utilized to train an XGBoost classifier. Then, all lncRNAs unrelated to the interested disease in experimental data were screened, each of lncRNA feature vectors was combined with the current disease feature vectors. Finally, all these lncRNA-disease feature pairs were inputted into the trained classifier, and the output scores were taken as the correlation probability between the lncRNAs and the corresponding disease. After sorting these scores by descending order, the top ten lncRNA-disease associations were selected. And the validity of selected associations was verified by searching the relative literature in the PubMed database. The results of case studies (Tables 3, 4, 5) and roughly analyses of each disease are as follows.

[Breast Cancer] According to the latest data of the global cancer burden in 2020 [41], there were 2.26 million new cases of Breast cancer worldwide in 2020, accounting for 11.7% of all new cases of cancer this year, ranking first among all cancers. Symptoms of breast cancer includes lumps in the breast, changes in the shape of the breast, depressions in the skin with bone pain, swollen lymph nodes, tachypnea or yellow skin. Table 3 shows the top-10 lncRNA-disease associations of unknown association of SVDNLDA for breast cancer prediction.

[Lung Cancer] Lung cancer is a kind of malignant lung tumor caused by uncontrolled cell growth in lung tissues, the malignant growth can spread beyond the lungs by metastasizing to nearby tissues or other parts of the body. In 2020, there were 2.2

Table 3 Case study results of breast neoplasms

Rank	lncRNA	PMID
1	HCP5	32165090
2	MBNL1-AS1	31113460
3	TNRC6C-AS1	30038597
4	RN7SL1	28709002
5	MIR155HG	32165090
6	DISC1	31783305
7	LRRC2-AS1	Unknown
8	NRON	32913541
9	lnc-KCTD6-3	Unknown
10	MORT	28690657

Table 4 Case study results of lung neoplasms

Rank	lncRNA	PMID
1	TNRC6C-AS1	32041817
2	SNHG16	33015794
3	TUG1	33073961
4	LINC00963	28923857
5	NRON	29772429
6	RN7SL1	<i>Unknown</i>
7	lnc-Sox5	23932921
8	HAGLR	28632999
9	LINC00460	32633366
10	AGAP2-AS1	32015683

Table 5 Case study results of pancreatic neoplasms

Rank	lncRNA	PMID
1	TNRC6C-AS1	32382761
2	FENDRR	33417179
3	FOXCT	<i>Unknown</i>
4	PCAT1	33629282
5	MBNL1-AS1	<i>Unknown</i>
6	lnc-KCTD6-3	32046932
7	MIR31HG	32134327
8	CASC9	33520364
9	DGCR5	33613108
10	MORT	26549028

million new cases of lung cancer worldwide, accounting for 11.4% of all the new cancer cases, ranking secondly among all cancers [41]. The most common symptoms of lung cancer include coughing, weight loss, breath hard and chest pain. Most of lung cancer cases are caused by long-term smoking. Table 4 illustrates the top-10 lncRNA results of lung cancer predicted by SVDNVLDA.

[Pancreatic Cancer] The common signs and symptoms of pancreatic cancer include yellow skin, abdominal or back pain, unexplained weight loss and loss of appetite. Usually, there are no obvious symptoms in the early stages of pancreatic cancer, yet when the symptoms are sufficient to indicate contraction generally means the disease is at an advanced stage, and by the time of diagnosis, pancreatic cancer has usually spread to other parts of the body. In the global statistics of cancer deaths in 2020, pancreatic cancer caused 466,000 deaths, and more than half of these clinical cases of pancreatic cancer were over 79 years old [41]. Table 5 presents the top-10 potential lncRNAs with pancreatic cancer prediction.

Among all the results of three diseases, the latest Pubmed literature support was found for 8, 9 and 8 of the top-10 predicted lncRNAs with maximum correlation probability, respectively. This clearly indicates that our model has a good performance in the prediction of actual disease-related lncRNAs, and possess potential application

value and scientific significance. Full results of the three cancers are given in Additional file 5, Additional file 6 and Additional file 7.

Discussions

In this paper, we proposed an integrated feature extraction model, SVDNVLDA, for predicting potential lncRNA-disease associations. In SVDNVLDA, the network representation learning method *node2vec* and matrix decomposition method SVD were originally integrated to predict the potential lncRNA-disease associations. It also can be regarded as an open framework, in which more feature extraction methods can be flexibly applied.

However, there are still some potential weaknesses in our model, which mainly relies on the limitations of the data used in this paper. Specifically, relying solely on the associated data almost could not comprehensively reflect the complex interactions between lncRNAs and the other biomolecules. Meanwhile, in the heterogeneous network LMDN, the node representations, obtained by *node2vec*, have been proven to be capable to retain the topology information of nodes in network, yet they fail to remain the information of different node types which is abundant and valuable in heterogeneous networks. It would be improved on the expansion of experimental data and introducing more advanced representation learning methods in future studies.

Conclusions

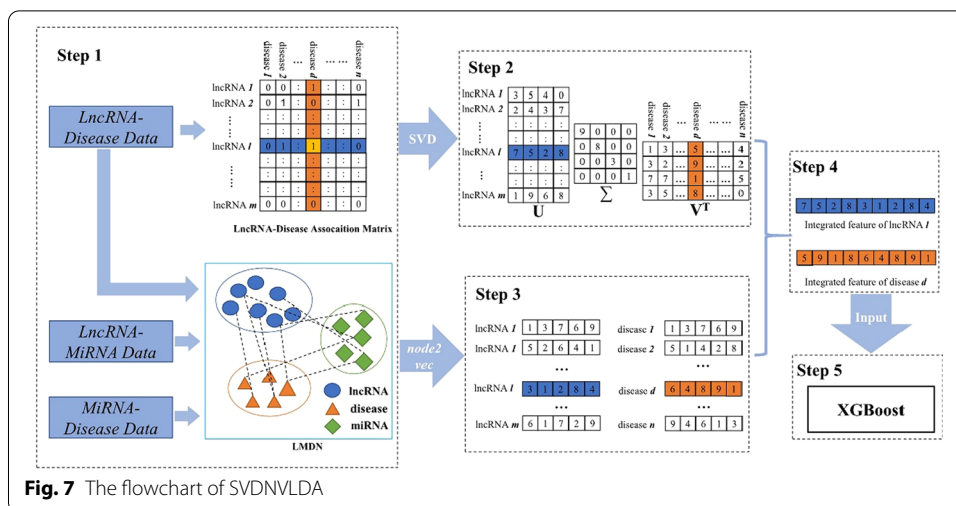
In SVDNVLDA, the linear feature representations of lncRNAs and diseases containing their linear interaction information were obtained by matrix decomposition method SVD; and the nonlinear features containing network topology information were obtained by *node2vec*. The integrated feature vectors of aforementioned features were inputted into a machine learning classifier, which transformed the lncRNA-disease association prediction into a binary classification problem. The AUC and AUPR values of SVDNVLDA are higher than any of five popular prediction methods under tenfold cross-validations. The prediction performance on data sets of different scales shows that SVDNVLDA can be adapted to a range of data sets and possess strong robustness. In addition, the case studies of three common cancers indicate its effectiveness in practical applications.

Materials and methods

Overview of SVDNVLDA

Matrix decomposition method, SVD, and network embedding method, *node2vec*, were novelly integrated in SVDNVLDA for obtaining the linear and the nonlinear representations of both lncRNA and disease entities respectively. By combining the different features of each lncRNA and each disease, the integrated feature vectors were constructed which fused the linear features of interaction information and the nonlinear features of network topology information. These feature vectors were served as the inputs of one machine learning classifier and the corresponding predicted results would be obtained in the end (Fig. 7).

Step1: Data processing and construction of lncRNA-disease association matrix and lncRNA-miRNA-disease association network (LMDN). Step 2: Apply SVD on association matrix to get linear features. Step 3: Apply *node2vec* on LMDN to get



nonlinear features. Step 4: Feature integration. Step 5: Use XGBoost classifier to predict association.

Data preprocessing

The study mainly included lncRNA-disease association data, lncRNA-miRNA association data and miRNA-disease association data. The experimentally confirmed lncRNA-disease association data were downloaded from LncRNADisease v2.0 [42] and Lnc2Cancer v3.0 [43]. All disease names were converted into standard MESH disease terms, and duplicate data was filtered to retain only one replication. For avoiding experimental errors that came from the downloaded data, the lncRNAs with one or none association were removed. In the end, a total number of 4518 associations between 861 lncRNAs and 253 diseases were obtained.

The known lncRNA-miRNA association data was downloaded from Encori [44] and NPInter V4.0 databases [45]. After eliminating redundancy, only records of the lncRNAs commonly to lncRNA-disease data and the miRNAs commonly to miRNA-disease data were selected. Finally, a total of 8172 lncRNA-miRNA associations were obtained involving 338 lncRNAs and 285 miRNAs.

As for miRNA-disease association data, it was obtained from the HMDD v3.2 database [46]. The original data includes two types of association records, namely the subjective causality and passive changes of miRNAs during the course of diseases. By contrast, the studies of miRNAs in causal relationship with diseases were more valuable for exploring the pathogenesis and searching for new biomarkers. In our experiment, only the related records with causal relationships in HMDD database were picked. All disease names were transformed to standardized names based on MeSH glossary, and the lncRNAs associated with only one disease were removed from the original data. Ultimately, a total count of 861 lncRNAs, 437 miRNAs and 431 diseases were involved in our experiment. The statistical overview of formed data, also as the statistical overview of LMDN was documented in Additional file 8.

Construct lncRNA-disease association matrix and LMDN

Firstly, the lncRNA- disease association matrix was constructed. For lncRNA l , if there is a known association with disease j in our collected data, the corresponding element value in the association matrix $R_{M \times N}$ is 1; otherwise, it is 0. The formula is made out as:

$$R_{M \times N}(i,j) = \begin{cases} 1, & \text{if } i \text{ and } j \text{ have association} \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

in our experiment, the real matrix $R_{M \times N}$ was shaped as 861×437 dimensions.

After the construction of association matrix, lncRNA-disease association data combined with lncRNA-miRNA association and miRNA-disease association data were used to construct lncRNA-miRNA-disease association heterogeneous network (LMDN). Among the three types of vertices in LMDN, namely lncRNA, miRNA and disease, there would be an edge between two vertices with association record, otherwise the two vertices would have no connection. The heterogeneous network was a sparse network with 1769 nodes and 16,878 edges, as detailed in Additional file 8.

Linear feature extraction based on singular value decomposition

SVD is a matrix decomposition method which has been widely used in recommender systems [47, 48]. In SVD, the matrix is common decomposed into the multiplying of three matrices:

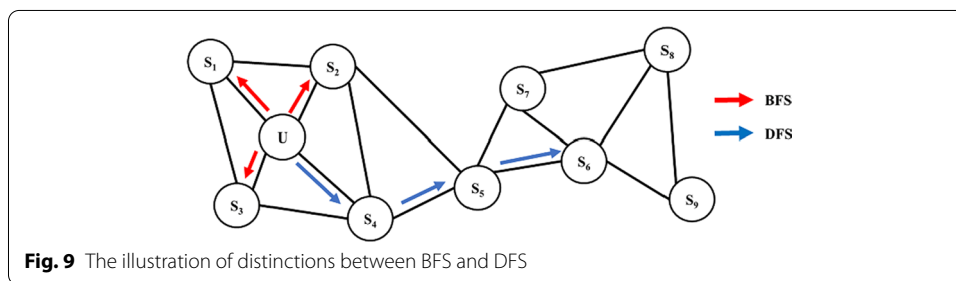
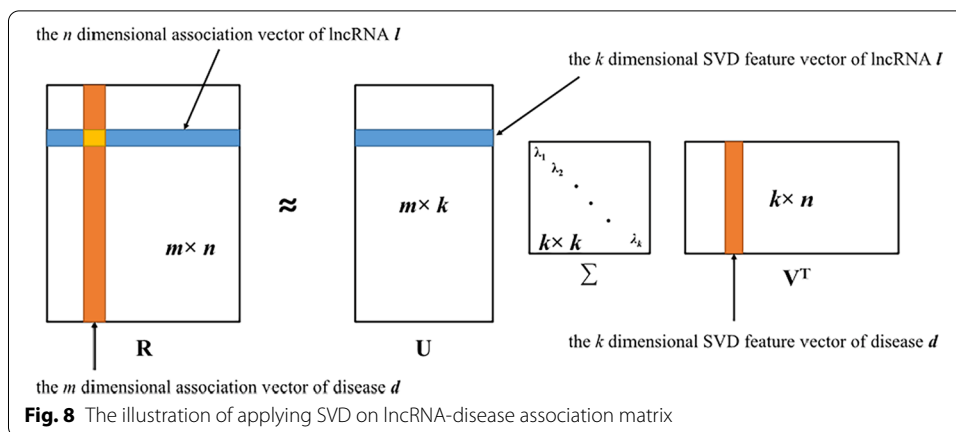
$$R_{M \times N} = U_{M \times C} \cdot \Sigma_{C \times C} \cdot V_{C \times N}^T \quad (2)$$

As a typical collaborative filtering-based recommendation system with SVD, the initial matrix R represents a rating matrix for M users' rates on N goods. Among the resulted matrixes, U represents the interesting levels of M users on C features of goods, namely users' characteristics or commodity affinity; while Σ represents the importance of each feature of goods, specified as a non-negative diagonal matrix, in which diagonal elements are arranged as descending order. V^T represents the distribution of C features in N goods [49].

Analogically, applying SVD on lncRNA-disease association matrix $R_{M \times N}$, the obtained matrixes U , Σ and V^T could represent lncRNA feature matrix, feature weight matrix and disease feature matrix, respectively. For dimensional reduction purpose, only the ranked k features with larger numerical values in Σ were taken, and R would be expressed as:

$$R_{M \times N} \approx U_{M \times k} \cdot \Sigma_{k \times k} \cdot V_{k \times N}^T \quad (3)$$

In fact, the binary matrix R is not an ideal initial matrix. In recommendation system, 0 (or blank) elements in rating matrixs cannot actually represent these rates of products, more likely, it is commonly due to missing users' evaluations. Thus, in lncRNA-disease association matrix R , the value 0 usually represents that the corresponding association has not been confirmed. Therefore, for calculation convenience and considering biological meaning, all the 0 elements in original binary matrix R

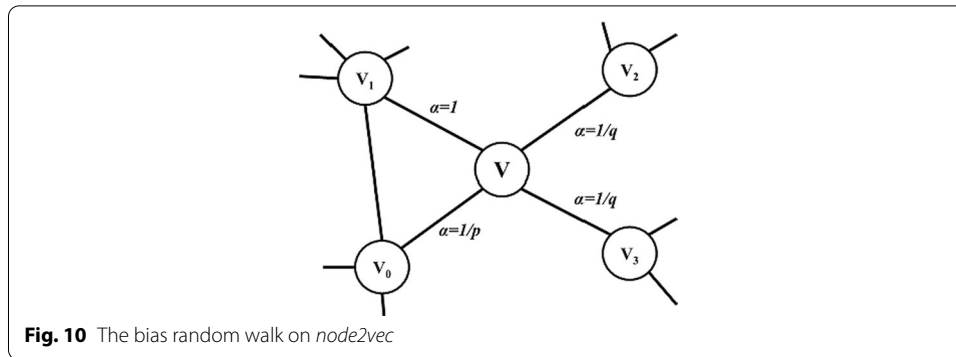


were replaced by 10^{-6} in our experiment. Based on the theory of SVD, each row in $U_{M \times k}$ represents a k -dimensional linear feature vector of a certain IncRNA. Similarly, each column in $V_{k \times N}^T$ represents a k -dimensional linear feature vector of a certain disease (Fig. 8).

Nonlinear feature extraction based on Node2vec

Network representation learning (NRL), also known as network embedding, refers to map nodes into a continuous low-dimensional space on the premise of keeping characteristics of nodes in the original network. Given a network $G = (V, E)$, where $V = \{v_i\}$ represents the collection of nodes and $E = e_i \subset \{V \times V\}$ represents the collection of edges. The mathematical expression of NRL is: $\forall v_i$, find a map $f : V \rightarrow R^d$, and $d \ll |V|$. The ideal learned node representations should be able to quantify the characteristics of nodes in social network, which could be intuitively expressed that topological neighbor nodes have small numerical vector distance and the representations of nodes in the same community have larger similarity than nodes outside the community. Up to now, many NRL methods have been widely used to solve problems such as node classification, community discovery, link prediction and data visualization [50].

As a semi-supervised network feature learning method, node2vec [51] innovatively proposed a biased random walk on the basis of word representation method [52] and DeepWalk [53], as well as defined a more flexible way to select the next step node with random walk. More specifically, node2vec trades off the two kinds of random walk strategy: Breadth-first search (BFS) and Depth-first search (DFS), which are shown in Fig. 9. Unlike the original random walk, node2vec can artificially



control the degree of BFS and DFS by adjusting parameters based on the preferences of actual practice scenario. Here is a detailed description of simple random walk and modified biased random walk in *node2vec* (Fig. 10).

For a given boot node u , simulate a simple unbiased random walk with l length. c_i represents the i^{th} node in the process of random walk. Let $c_0 = u$, and the transition probability of the node reached in i^{th} step is:

$$P(c_i = x | c_{i-1} = v) = \begin{cases} \frac{\pi_{vx}}{Z}, & \text{if } (v,x) \in E \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

of which π_{vx} is the unnormalized transition probability between nodes v and x , Z represents a normalized constant term.

As for the biased random walk in *node2vec*, just as shown in Fig. 10, if the root position of a random walk is set at node t , through edge (t, v) , the current position reached node v , and the transition probability is set as follows:

$$\alpha_{pq}(t, x) = \begin{cases} \frac{1}{p} & \text{if } d_{tx} = 0 \\ 1 & \text{if } d_{tx} = 1 \\ \frac{1}{q} & \text{if } d_{tx} = 2 \end{cases} \tag{5}$$

d_{tx} represents the shortest distance between nodes t and x and the possible value of d_{tx} is 0,1,2. As shown in Fig. 10, the parameter p controls the probability that the next step of walk will return to the previous node. If p is greater than 1, the random walk will have less tendency to turn back. The value of q controls the preference of BFS and DFS to guide the bias of random walk. If q is greater than 1, the random walk will be more inclined to BFS, that is, to the neighbor node of the starting node. If q is less than 1, the random walk is more inclined to DFS, that is, to go away from the starting node. When the values of p and q are both equal to 1, *node2vec* is equal to DeepWalk.

In the constructed LMDN, *node2vec* was adopted to obtain the corresponding representations for vertices. The representations of lncRNA and disease nodes generated by *node2vec* retain the topological information of the nodes in LMDN. The experimental results demonstrate that the obtained nonlinear features could effectively enhance the SVD based linear features and improve the information richness in integrated features.

Feature integration

Based on the decomposition of $R_{M \times N}$ and NRL method node2vec, we have obtained the linear feature matrixes U , V^T , and the nonlinear feature representations of lncRNA and disease nodes in LMDN. For each lncRNA i and disease j , the feature integration rules are as follows:

The linear features corresponding to lncRNA i is the i th row of U , which is noted as LL_i after being converted into a column vector. Similarly, the linear features corresponding to disease j is the j th column of V^T , represented as LD_j . The nonlinear features corresponding to i is noted as NL_i as well as the nonlinear features corresponding to j is noted as ND_j . The final integrated features of i and j is expressed as:

$$FL_i = \begin{bmatrix} LL_i \\ NL_i \end{bmatrix} \quad (6)$$

$$FD_j = \begin{bmatrix} LD_j \\ ND_j \end{bmatrix} \quad (7)$$

where $\begin{bmatrix} \end{bmatrix}$ represents the vector connect operation.

Abbreviations

SVD: Singular Value Decomposition; ncRNAs: Non-coding RNAs; RWR: Random walk with restart; LRLSLDAL: Laplacian Regularized Least Squares for lncRNA-Disease Association; SVDNLDA: Singular Value Decomposition and Node2Vec based lncRNA-Disease Association prediction model; Acc: Accuracy; Sen: Sensitivity; Spec: Specificity; Prec: Precision; MCC: Matthews correlation coefficient; LR: Linear regression; NB: Naïve Bayes; RF: Random Forest; ADB: AdaBoost; XGB: XGBoost; LMDN: lncRNA-miRNA-disease interaction heterogeneous network; NRL: Network representation learning; BFS: Breadth-first search; DFS: Depth-first search.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12859-021-04457-1>.

Additional file 1. The numerical values of diagonal elements in the importance matrix Σ .

Additional file 2. The sensitivity results of different features in classifiers.

Additional file 3. The specificity results of different features in classifiers.

Additional file 4. The precision results of different features in classifiers.

Additional file 5. The complete case study results of breast cancer.

Additional file 6. The complete case study results of lung cancer.

Additional file 7. The complete case study results of pancreatic cancer.

Additional file 8. The statistical overview of experimental data.

Acknowledgements

Not applicable.

Authors' contributions

JS and KF conceived, led the project, evaluated the methods, suggested improvements and analyzed the results. JL (Jianwei) and JL (Jianing) conducted the experiments and wrote the manuscript. MK and DW collected, organized data and modified the manuscript. All authors read and approved the final manuscript.

Funding

This study was supported by National Natural Science Foundation of China (81672113, 62072154) and Natural Science Foundation of Hebei Province (C2018202083).

Availability of data and materials

The source code and data of SVDNLDA are available at <https://github.com/iALKing/SVDNLDA>.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent to publish

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Institute of Computational Medicine, School of Artificial Intelligence, Hebei University of Technology, Tianjin 300401, China. ²Hebei Province Key Laboratory of Big Data Calculation, Hebei University of Technology, Tianjin 300401, China. ³School of Life Sciences, Tiangong University, Tianjin 300387, China.

Received: 2 September 2021 Accepted: 18 October 2021

Published online: 02 November 2021

References

- Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*. 2009;458(7235):223–7.
- Xue M, Zhuo Y, Shan B. MicroRNAs, long noncoding RNAs, and their functions in human disease. *Methods Mol Biol*. 2017;1617:1–25.
- DiStefano JK. The emerging role of long noncoding RNAs in human disease. *Methods Mol Biol*. 2018;1706:91–110.
- Chen J, Shishkin AA, Zhu X, Kadri S, Maza I, Guttman M, Hanna JH, Regev A, Garber M. Evolutionary analysis across mammals reveals distinct classes of long non-coding RNAs. *Genome Biol*. 2016;17(1):1–17.
- McDonel P, Guttman M. Approaches for understanding the mechanisms of long noncoding RNA regulation of gene expression. *Cold Spring Harb Perspect Biol*. 2019;11(12):a032151.
- Tsang W, Kwok T. Riboregulator H19 induction of MDR1-associated drug resistance in human hepatocellular carcinoma cells. *Oncogene*. 2007;26(33):4877–81.
- Li Y, Zhuang L, Wang Y, Hu Y, Wu Y, Wang D, Xu J. Connect the dots: a systems level approach for analyzing the miRNA-mediated cell death network. *Autophagy*. 2013;9(3):436–9.
- Munos B. Lessons from 60 years of pharmaceutical innovation. *Nat Rev Drug Discov*. 2009;8(12):959–68.
- Lalève S, Feil R. Long noncoding RNAs in human disease: emerging mechanisms and therapeutic strategies. *Epigenomics*. 2015;7(6):877–9.
- Mercer TR, Qureshi IA, Gokhan S, Dinger ME, Li G, Mattick JS, Mehler MF. Long noncoding RNAs in neuronal-glia fate specification and oligodendrocyte lineage maturation. *BMC Neurosci*. 2010;11(1):1–15.
- Mercer TR, Mattick JS. Structure and function of long noncoding RNAs in epigenetic regulation. *Nat Struct Mol Biol*. 2013;20(3):300–7.
- Quinodoz S, Guttman M. Long noncoding RNAs: an emerging link between gene regulation and nuclear organization. *Trends Cell Biol*. 2014;24(11):651–63.
- Bhan A, Mandal SS. Long noncoding RNAs: emerging stars in gene regulation, epigenetics and human disease. *ChemMedChem*. 2014;9(9):1932–56.
- Engreitz JM, Ollikainen N, Guttman M. Long non-coding RNAs: spatial amplifiers that control nuclear structure and gene expression. *Nat Rev Mol Cell Biol*. 2016;17(12):756–70.
- Gutschner T, Hämmerle M, Eißmann M, Hsu J, Kim Y, Hung G, Revenko A, Arun G, Stentrup M, Groß M. The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells. *Can Res*. 2013;73(3):1180–9.
- Topel H, Bagirsakci E, Comez D, Bagci G, Cakan-Akdogan G, Atabey N. lncRNA HOTAIR overexpression induced downregulation of c-Met signaling promotes hybrid epithelial/mesenchymal phenotype in hepatocellular carcinoma cells. *Cell Commun Signal*. 2020;18(1):1–19.
- Chen X, Yan G-Y. Novel human lncRNA–disease association inference based on lncRNA expression profiles. *Bioinformatics*. 2013;29(20):2617–24.
- Sun J, Shi H, Wang Z, Zhang C, Liu L, Wang L, He W, Hao D, Liu S, Zhou M. Inferring novel lncRNA–disease associations based on a random walk model of a lncRNA functional similarity network. *Mol Biosyst*. 2014;10(8):2074–81.
- Yao Q, Wu L, Li J, Guang Yang L, Sun Y, Li Z, He S, Feng F, Li H, Li Y. Global prioritizing disease candidate lncRNAs via a multi-level composite network. *Sci Rep*. 2017;7(1):1–13.
- Ding L, Wang M, Sun D, Li A. TPGLDA: Novel prediction of associations between lncRNAs and diseases via lncRNA–disease–gene tripartite graph. *Sci Rep*. 2018;8(1):1–11.
- Zhao X, Yang Y, Yin M. MHRWR: prediction of lncRNA–disease associations based on multiple heterogeneous networks. *IEEE/ACM Trans Comput Biol Bioinforma*. 2020;PP(99):1–1.
- Xie G, Jiang J, Sun Y. LDA-LNSUBRW: lncRNA–disease association prediction based on linear neighborhood similarity and unbalanced bi-random walk. *IEEE/ACM Trans Comput Biol Bioinform*. 2020;PP(99):1–1.
- Xie G, Huang Z, Liu Z, Lin Z, Ma L. NCPHLDA: a novel method for human lncRNA–disease association prediction based on network consistency projection. *Mol Omics*. 2019;15(6):442–50.
- Liu M-X, Chen X, Chen G, Cui Q-H, Yan G-Y. A computational framework to infer human disease-associated long noncoding RNAs. *PLoS ONE*. 2014;9(1):e84408.
- Guo Z-H, You Z-H, Wang Y-B, Yi H-C, Chen Z-H. A learning-based method for lncRNA–disease association identification combining similarity information and rotation forest. *IScience*. 2019;19:786–95.

26. Zeng M, Lu C, Zhang F, Lu Z, Wu F-X, Li Y, Li M. LncRNA–disease association prediction through combining linear and non-linear features with matrix factorization and deep learning techniques. In: 2019 IEEE international conference on bioinformatics and biomedicine (BIBM). IEEE; 2019. pp. 577–582.
27. Zeng M, Lu C, Fei Z, Wu F, Li Y, Wang J, Li M. DMFLDA: a deep learning framework for predicting lncRNA-disease associations. *IEEE/ACM Trans Comput Biol Bioinform*. 2020;PP(99):1–1.
28. Zeng M, Lu C, Zhang F, Li Y, Wu F-X, Li Y, Li M. SDLDA: lncRNA-disease association prediction based on singular value decomposition and deep learning. *Methods*. 2020;179:73–80.
29. Chen X. Predicting lncRNA-disease associations and constructing lncRNA functional similarity network based on the information of miRNA. *Sci Rep*. 2015;5(1):1–11.
30. Lu C, Yang M, Luo F, Wu F-X, Li M, Pan Y, Li Y, Wang J. Prediction of lncRNA-disease associations based on inductive matrix completion. *Bioinformatics*. 2018;34(19):3357–64.
31. Fan Y, Chen M, Pan X. GCRFLDA: scoring lncRNA-disease associations using graph convolution matrix completion with conditional random field. *Brief Bioinform*. 2021;22:438–450.
32. Boughorbel S, Jarray F, El-Anbari M. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLoS ONE*. 2017;12(6):e0177678.
33. Chicco D, Töttsch N, Jurman G. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData Min*. 2021;14(1):1–22.
34. Hanczar B, Hua J, Sima C, Weinstein J, Bittner M, Dougherty ER. Small-sample precision of ROC-related estimates. *Bioinformatics*. 2010;26(6):822–30.
35. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*. 2015;10(3):e0118432.
36. Rish I. An empirical study of the naive Bayes classifier. In: *IJCAI 2001 workshop on empirical methods in artificial intelligence*. 2001. pp. 41–46.
37. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5–32.
38. Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. *J Comput Syst Sci*. 1997;55(1):119–39.
39. Chen T, Guestrin C. Xgboost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016. pp. 785–794.
40. Zhang H, Liang Y, Peng C, Han S, Du W, Li Y. Predicting lncRNA-disease associations using network topological similarity based on deep mining heterogeneous networks. *Math Biosci*. 2019;315:108229.
41. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2021;71(3):209–49.
42. Bao Z, Yang Z, Huang Z, Zhou Y, Cui Q, Dong D. LncRNADisease 2.0: an updated database of long non-coding RNA-associated diseases. *Nucleic Acids Res*. 2019;47(D1):D1034–7.
43. Gao Y, Shang S, Guo S, Li X, Zhou H, Liu H, Sun Y, Wang J, Wang P, Zhi H. Lnc2Cancer 3.0: an updated resource for experimentally supported lncRNA/circRNA cancer associations and web tools based on RNA-seq and scRNA-seq data. *Nucleic Acids Res*. 2021;49(D1):D1251–8.
44. Li J-H, Liu S, Zhou H, Qu L-H, Yang J-H. starBase v.20: decoding miRNA-ceRNA, miRNA-ncRNA and protein–RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res*. 2014;42(D1):D92–7.
45. Teng X, Chen X, Xue H, Tang Y, Zhang P, Kang Q, Hao Y, Chen R, Zhao Y, He S. NPInter v4.0: an integrated database of ncRNA interactions. *Nucleic Acids Res*. 2020;48(D1):D160–5.
46. Huang Z, Shi J, Gao Y, Cui C, Zhang S, Li J, Zhou Y, Cui Q. HMDD v3.0: a database for experimentally supported human microRNA–disease associations. *Nucleic Acids Res*. 2019;47(D1):D1013–7.
47. Vozalis MG, Margaritis KG. Applying SVD on item-based filtering. In: *5th international conference on intelligent systems design and applications (ISDA'05)*. IEEE. 2005. pp. 464–469.
48. Vozalis MG, Margaritis KG. Using SVD and demographic data for the enhancement of generalized collaborative filtering. *Inf Sci*. 2007;177(15):3017–37.
49. Cheng W, Yin G, Dong Y, Dong H, Zhang W. Collaborative filtering recommendation on users' interest sequences. *PLoS ONE*. 2016;11(5):e0155739.
50. Yang C, Sun M, Liu Z, Tu C. Fast network embedding enhancement via high order proximity approximation. In: *IJCAI: 2017*. pp. 3894–3900.
51. Grover A, Leskovec J. node2vec: scalable feature learning for networks. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016. pp. 855–864.
52. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. 2013. arXiv:13013781
53. Perozzi B, Al-Rfou R, Skiena S. Deepwalk: online learning of social representations. In: *Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining*. 2014. pp. 701–710.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.