A Novel Computational Model for Predicting Potential LncRNA-Disease Associations based on Both Direct and Indirect Features of LncRNA-Disease Pairs

Yubin Xiao[1,3], Zheng Xiao[2], Xiang Feng[1], Zhiping Chen[1], Linai Kuang[3], Lei Wang[*1,3]

[1]College of Computer Engineering & Applied Mathematics, Changsha University, Changsha 410001, PR China

[2] Hunan Province Key Laboratory of Tumor Cellular & Molecular Pathology, Cancer Research Institute, University of South China, Hengyang, Hunan 421001, PR China.

[3] Key Laboratory of Hunan Province for Internet of Things and Information Security, Xiangtan University, Xiangtan 411105, PR China

Yubin Xiao, is with the College of Information Engineering, Xiangtan University, Xiangtan, Hunan, China(email: xiaoyb@smail.xtu.edu.cn)

Zheng Xiao, is with the University of South China, Hengyang, Hunan, China(email: 296926363@qq.com)

Xiang Feng, is with the College of Engineering & Applied Mathematics, Changsha University, Changsha, Hunan, China(email: fengxiang@xtu.edu.cn)

Zhiping Chen, is with the College of Engineering & Applied Mathematics, Changsha University, Changsha, Hunan, China(email: zpchen@ccsu.edu.cn)

Linai Kuang, is with the College of Information Engineering, Xiangtan University, Xiangtan, Hunan, China(email: kla@xtu.edu.cn)

Lei Wang, is with the College of Information Engineering, Xiangtan University, Xiangtan, Hunan, China(email: wanglei@xtu.edu.cn)

*Corresponding Author, Lei Wang, is with the College of Information Engineering, Xiangtan University, Xiangtan, Hunan, China(email: wanglei@xtu.edu.cn)

Abstract

Background: Accumulating evidence has demonstrated that long non-coding RNAs (lncRNAs) are closely associated with human diseases, and it is helpful for the diagnosis and treatment of diseases to get the relationships between lncRNAs and diseases. Due to the high costs and time complexity of traditional bio-experiments, in recent years, more and more computational methods

have been proposed by researchers to infer potential lncRNA-disease associations. However, there exist all kinds of limitations in these state-of-the-art prediction methods as well.

Results: In this manuscript, a novel computational model named FVTLDA is proposed to infer potential lncRNA-disease associations. In FVTLDA, its major novelty lies in the integration of direct and indirect features related to lncRNA-disease associations such as the feature vectors of lncRNA-disease pairs and their corresponding association probability fractions, which guarantees that FVTLDA can be utilized to predict diseases without known related-lncRNAs and lncRNAs without known related-diseases. Moreover, FVTLDA neither relies solely on known lncRNA-disease nor requires any negative samples, which guarantee that it can infer potential lncRNA-disease associations more equitably and effectively than traditional state-of-the-art prediction methods. Additionally, to avoid the limitations of single model prediction techniques, we combine FVTLDA with the Multiple Linear Regression (MLR) and the Artificial Neural Network (ANN) for data analysis respectively. Simulation experiment results show that FVTLDA with MLR can achieve reliable AUCs of 0.8909, 0.8936 and 0.8970 in 5-Fold Cross Validation (5-fold CV), 10-Fold Cross Validation (10-fold CV) and Leave-One-Out Cross Validation (LOOCV), separately, while FVTLDA with ANN can achieve reliable AUCs of 0.8766, 0.8830 and 0.8807 in 5-fold CV, 10-fold CV, and LOOCV respectively. Furthermore, in case studies of gastric cancer, leukemia and lung cancer, experiment results show that there are 8, 8 and 8 out of top 10 candidate lncRNAs predicted by FVTLDA with MLR, and 8, 7 and 8 out of top 10 candidate lncRNAs predicted by FVTLDA with ANN, having been verified by recent literature. Moreover, comparing with the representative prediction model of KATZLDA, comparison results illustrate that FVTLDA with MLR and FVTLDA with ANN can achieve the average case study contrast scores of 0.8429 and 0.8515 respectively, which are both notably higher than the average case study contrast score of 0.6375 achieved by KATZLDA.

Keywords: LncRNA-disease association prediction; features; Random walk; Multiple Linear Regression; Artificial Neural Network.

Background

LncRNAs have long been considered as a transcriptional noise [1-2]. However, in recent years, more and more researches have shown that lncRNAs play key roles in numerous important biological processes of humans, including chromatin modification, epigenetic regulation, cell cycle control, cell differentiation and so on [3–5]. Especially, accumulating bio-experiments have confirmed that mutations and dysregulations of lncRNAs are associated with the development of diseases, such as leukemia [6], neurological disorders [7], coronary artery diseases [8] and several cancers [9]. Hence, effectively inferring potential associations between lncRNAs and diseases can not only help understand the pathogenesis of some complex diseases at the molecular level, but also be conducive to provide biomarkers for disease diagnosis, therapy and prognosis. Up to now, along with the rapid increasement of newly inferred lncRNAs, some publicly available lncRNA-related databases, including lncRNADisease [10], NONCODE [11], lncRNAdb [12] and NRED [13], have been established successively. However, the number of known lncRNA-disease

associations is still very limited, since traditional biological experiments are costly and time-consuming. Therefore, it is important and necessary to construct effective and high-throughput computational models to explore potential lncRNA-disease associations.

So far, researchers have developed numerous powerful computational models to predict potential lncRNA-disease associations, which can be roughly classified into three major categories according to their main implementation strategies [14]. Among them, the first category aims to adopt machine learning methods to predict potential lncRNA-disease associations. For example, Yu and Wang et al. proposed a prediction model based on the Naïve Bayes classifier [15] in 2018 and a prediction model based on the collaborative filtering algorithm [16] in 2019 to infer potential lncRNA-disease associations respectively. Xuan and Wang et al. developed a probabilistic matrix factorization model based on the semi-supervised learning method to identify potential associations between lncRNAs and diseases [17]. In these prediction models of the first category, the major drawback lies in the requirement of negative samples as the training set, which will affect their prediction performances notably, since the negative samples are usually difficult to obtain.

Different from the first category, the second category focuses on implementing propagation algorithms such as Random Walk on a heterogeneous network constructed by integrating lncRNA-disease association network, disease similarity network and lncRNA similar network, etc. For instance, in 2014, sun et al. established a global network-based computational model, which adopted the random walk with restart (RWR) algorithm to predict potential lncRNA-disease associations [18]. In 2015, Zhou et al. proposed a prediction model by implementing RWR on a heterogeneous network consisting of known lncRNA-disease association network, miRNA-associated lncRNA crosstalk network and disease similarity network [19]. However, these two models mentioned above can only be applied to infer lncRNAs with related-disease or known miRNA-disease associations. To break through this kind of limitation, in 2015, Chen et al. developed a computational model called KATZLDA for prediction of potential lncRNA-disease associations [20], which can infer potential lncRNAs in the absence of known associated diseases. But it may cause the bias to lncRNAs with more known related-diseases and diseases with more known related-lncRNAs as well due to its construction of the network.

According to the above descriptions, it is obvious that the prediction performance of all these models of both categories will be greatly influenced by the number of known lncRNA-disease associations. However, the number of known lncRNA-disease associations confirmed by bio-experiments is still very limited. Therefore, to avoid the drawback of limited known lncRNA-disease associations, the third category adopt indirect biological information to explore the prediction of potential lncRNA-disease associations. For instance, in 2014, Liu et al. proposed a novel prediction model by combining human lncRNA expression profiles, human disease-associated gene data and gene expression profiles [21], which can achieve exciting prediction performance while there are no known lncRNA-disease associations. However, it cannot be implemented to predict lncRNAs without gene-related records.

Different from the above existing methods, in this manuscript, a novel computational model named FVTLDA is proposed to reveal potential lncRNA-disease associations. In FVTLDA, to avoid the requirement of negative samples and improve the prediction accuracy of FVTLDA while there is a lack of known associations between lncRNAs and diseases, both direct and

indirect biological information including known lncRNA-disease associations, known miRNA-disease associations and known miRNA-lncRNA associations are introduced first. And among them, known lncRNA-disease associations will be utilized to extract direct features called Association Probability Fractions for lncRNA-disease pairs based on the concept of Disease Clique. Meanwhile, indirect biological information including known miRNA-disease associations and known miRNA-lncRNA associations will be utilized to extract indirect features called feature vectors for lncRNA-disease pairs by adopting the random walk with restart. And then, to avoid the limitation of single model prediction techniques, based on the direct and indirect features obtained for lncRNA-disease pairs, the Multiple Linear Regression (MLR) and Artificial Neural Network (ANN) will be combined with FVTLDA to reveal potential lncRNA-disease associations respectively. Furthermore, to estimate the prediction performance of FVTLDA, different frameworks including the LOOCV, 5-fold CV and 10-fold CV are implemented to compare FVTLDA with existing competing models. Simulation experiment results show that FVTLDA with MLR can achieve reliable AUCs of 0.8909, 0.8936 and 0.8970 in 5-fold CV, 10-fold CV and LOOCV respectively, while FVTLDA with ANN can achieve reliable AUCs of 0.8766, 0.8830 and 0.8807 in 5-fold CV, 10-fold CV and LOOCV separately, which both outperform existing state-of-the-art models. Meanwhile, in case studies of gastric cancer, leukemia and lung cancer, simulation experiment results show that there are 8, 8 and 8 out of top 10 candidate lncRNAs predicted by FVTLDA with MLR, and 8, 7 and 8 out of top 10 candidate lncRNAs predicted by FVTLDA with ANN, having been verified respectively in biological experimental studies or other independent studies. Finally, in order to further illustrate the actual predictive ability of FVTLDA, we have compared it with the representative prediction model KATZLDA based on the new concept of case study contrast score as well. And simulation experiment results show that the average case study contrast scores of FVTLDA with MLR and FVTLDA with ANN are 0.8429 and 0.8515 respectively, which both outperform the average case study contrast score of 0.6375 obtained by KATZLDA notably.

RESULT

Performance evaluation

In order to evaluate the prediction performance of FVTLDA, in this section, we implement the LOOCV on FVTLDA as follows: For all known lncRNA-disease pairs obtained above, each pair with known correlations will be left out in turn for testing, and other lncRNA-disease pairs will retain as training samples for model learning. Particularly, testing samples and lncRNA-disease pairs without known correlations will be considered as candidates. After the implementation of FVTLDA, the ranking positions of test samples in candidates can be obtained according to the association probability fractions. If the ranking of a test sample is above the given threshold, it will be seen as a successful prediction or a positive sample. Otherwise, it is seen as an unsuccessful prediction or a negative sample. Besides, upon different thresholds, the corresponding True Positive Rate (TPR, sensitivity) and False Positive Rate (FPR, 1-specificity) can be calculated as follows:

$$TPR = \frac{TP}{TP + FN} \quad (1)$$

$$FPR = \frac{FP}{TN + FP} \quad (2)$$

Here, TP and TN represent the correctly identified positive and negative samples separately, while FP and FN denote the incorrectly identified positive and negative samples respectively.

Based on the above equations, the Receiver Operating Characteristic (ROC) curve can be drawn according to the TPRs and FPRs of different thresholds, and the area under ROC curve (AUC) will further be calculated to evaluate the performance of FVTLDA. The AUC value of 1 indicates the perfect prediction performance while the AUC value of 0.5 means random guess.

During simulation, we first compared FVTLDA_MLR (i.e., FVTLDA with MLR) with six state-of-the-art prediction models such as NBCLDA [15], CFNBC [16], PMFILDA [17], KATZLDA [20], SIMCLDA [22] and IIRWR [39] in the framework of LOOCV, and comparison results were shown in Figure.1. Through observing the Figure.1, it is easy to see that FVTLDA_MLR can achieve a reliable AUC of 0.8970, which significantly outperforms those six state-of-the-art prediction models with the increasement of AUC values by at least 0.0311.

Moreover, to eliminate the random error caused by the random initialization of weights and biases in FVTLDA_ANN (i.e., FVTLDA with ANN), during simulation, we repeat executing LOOCV on FVTLDA ANN for 20 times, and take the mean and variance of the AUC values as the final result. As illustrated in the following Figure.2, it is easy to see that FVTLDA ANN can achieve a reliable mean of AUC value of 0.8807 and standard deviation (std) of 0.0047 in LOOCV, which outperform these six state-of-the-art prediction models as well.

In order to further verify the prediction performance of FVTLDA while there are few known lncRNA-disease associations, the frameworks of *K*-fold CV including 5-fold CV and 10-fold CV are implemented to compare FVTLDA_MLR with other representative prediction models. During implementing the *K*-fold CV, all known lncRNA-disease associations are equally divided into *K* parts, each part will be left out as the test sample in turn, and other remaining lncRNA-disease pairs will be used as the training samples. As shown in the following Figure.3 and Figure.4, it is obvious that FVTLDA_MLR can achieve better predictive performance than the other six competing models, which demonstrates that FVTLDA can perform excellently in sparse data sets as well.

Furthermore, in order to eliminate the effects of the random partition of training samples, during simulation, we repeat the implementations of 5-CV and 10-CV several times respectively, and take the mean and variance of AUC value as the final results. As shown in the following Figure.5 and Figure.6, it is easy to find that FVTLDA_MLR can achieve the mean AUC value of 0.8903 and std of 0.0022 in 5-CV, and the mean AUC of 0.8940 and std of 0.0014 in 10-CV separately. Meanwhile, as for FVTLDA_ANN, from observing the following Figure.7 and Figure.8, it is also easy to see that it can achieve the mean AUC value of 0.8766 and std of 0.0043 in 5-CV, and the mean AUC of 0.8830 and std of 0.0022 in 10-CV respectively.

Finally, in order to demonstrate that FVTLDA can perform well in different data sets, we further compare it with other state-of-the-art models including HGLDA [23] and the method proposed by Yang et al. [24] in the framework of LOOCV. While comparing FVTLDA with HGLDA, we adopt the data set given by HGLDA, which consists of 183 experimentally validated

lncRNA-disease associations. While comparing FVTLDA with the method proposed by Yang et al., we adopt the dataset put forward by Yang et al., which consists of 319 known lncRNA-disease associations. As illustrated in the following Figure.9 and Figure.10, it is easy to see that FVTLDA outperforms these two kinds of models in different datasets.

Parameter analysis

In this section, the influences of parameters in FVTLDA will be estimated. From above descriptions, it is easy to know that the parameters r_1 and r_2 in Eq (11) and Eq (14) represent the restart probabilities of the random walk, the parameter *rate* in Eq (19) stands for the adjustment factor, and the parameters k_1 and k_2 in Eq (20) and Eq (21) denote the attenuation factors respectively.

In order to determine the optimal values of the above five parameters efficiently, we traverse the approximate range of each parameter with precision 0.1 through FVTLDA with MLR in the framework of LOOCV. For parameters that can further improve the precision, we take the approximate solution of the previous step as the default value, and then, the optimal solution with higher precision is obtained through traversal. As illustrated in the following Table 1, Table 2, Table 3, Table 4 and Table 5, it is obvious that the optimal values for these five parameters such as *rate*, r_1, r_2, k_1, and k_2 are 0.3, 0.001, 0.001, 0.008, 0.007 separately.

Table 1: Effects of the parameter rate to the performance of FVTLDA_MLR in LOOCV

rate	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
AUC	0.89698	0.89699	0.89699	**0.89701**	0.89697	0.89695	0.89696	0.89697	0.89695	0.89694	0.89693

Table 2: Effects of the parameter *r*1 to the performance of FVTLDA_MLR in LOOCV

r_1	0	0.001	0.002	0.003	0.004	0.005	0.006	0.007	0.008	0.009	0.01
AUC	0.72933	**0.89701**	0.89694	0.89689	0.89678	0.89673	0.89671	0.89668	0.89665	0.89660	0.89647

Table 3: Effects of the parameter *r2* to the performance of FVTLDA_MLR in LOOCV

r_2	0	0.001	0.002	0.003	0.004	0.005	0.006	0.007	0.008	0.009	0.01
AUC	0.79486	**0.89701**	0.89700	0.89692	0.89692	0.89693	0.89692	0.89693	0.89684	0.89680	0.89647

Table 4: Effects of the parameter *k*1to the performance of FVTLDA_MLR in LOOCV

k_1	0	0.001	0.002	0.003	0.004	0.005	0.006	0.007	0.008	0.009	0.01
AUC	0.89699	**0.89701**	0.89700	0.89700	0.89698	0.89695	0.89698	0.896700	0.89701	0.89700	0.89699

Table 5: Effects of the parameter *k2* to the performance of FVTLDA_MLR in LOOCV

k_2	0	0.001	0.002	0.003	0.004	0.005	0.006	0.007	0.008	0.009	0.01
AUC	0.89698	0.89699	0.89700	0.89699	0.89700	0.89700	0.89700	**0.89701**	0.89699	0.89697	0.89698

Case study

In order to further demonstrate the predictive ability of FVTLDA, in this section, we select gastric cancer, leukemia and lung cancer as case studies. During the simulation, for any given disease $d_i \in$ {the gastric cancer, the leukemia, the lung cancer}, only those lncRNAs that do not have known associations with d_i will be considered as validated candidates for d_i. Next, all candidate lncRNAs will be ranked according to their association probability fractions calculated by FVTLDA. Finally, the top 10 candidate d_i-related lncRNAs will be verified by recent articles and experiments published in the NCBI database (https://www.ncbi.nlm.nih.gov/). Additionally, in order to compare the difference of prediction performance between FVTLDA_MLR and FVTLDA_ANN, as well as the difference of prediction performance between FVTLDA and another representative prediction model KATZLDA, we further list all these lncRNAs in the top 10 candidate d_i-related lncRNAs predicted by FVTLDA_MLR, FVTLDA_ANN and KATZLDA separately. Simultaneously, we will provide corresponding rankings and relevant evidence of these lncRNAs as well. Moreover, in order to visualize the predictive ability of these three kinds of prediction models in the above case studies, we propose a novel concept of case study contrast score, which can be calculated as follows:

$$score = \exp\left(-\sum_{i=1}^{m}\frac{1}{i}-\frac{1}{R_i}\right) \qquad (3)$$

Here, m denotes the number of verified lncRNAs in top 10 predicted candidate lncRNAs, and R_i represents the ranking corresponding to the i^{th} confirmed lncRNA.

Gastric cancer is the second leading cause of cancer death [25]. Up to now, there are a large number of lncRNAs having been proved to be related to gastric cancer [26-27]. As shown in the following Table 6, it is obvious that FVTLDA_MLR, FVTLDA_ANN and KATZLDA can successfully predict 8, 8 and 8 confirmed lncRNAs out of the top 10 candidate lncRNAs respectively, and their corresponding case study contrast scores are 0.7168, 0.8377 and 0.8439 separately.

Table 6: Top 10 potential gastric cancer-related lncRNAs and their PubMed unique identifiers predicted by FVTLDA_MLR, FVTLDA_ANN and KATZLDA

disease	lncRNA	Evidence(PMID)	Rank by FVTLDA_MLR	Rank by FVTLDA_ANN	Rank by KATZLDA
gastric cancer	MALAT1	29719612	1	1	1
gastric cancer	PVT1	25258543	2	3	2
gastric cancer	HOTAIRM1	unknown	3	11	15
gastric cancer	GAS5	29557411	4	2	3
gastric cancer	TUG1	29719612	5	4	4
gastric cancer	NEAT1	28401449	6	7	5
gastric cancer	XIST	27620004	7	5	6
gastric cancer	KCNQ1OT1	unknown	8	6	8

gastric cancer	HOXA11-AS	28441948	9	13	27
gastric cancer	MIAT	29540201	10	10	7
gastric cancer	SNHG7	29131253	29	8	34
gastric cancer	TP73-AS1	unknown	17	9	35
gastric cancer	ZNRD1-AS1	unknown	12	60	9
gastric cancer	HOTTIP	27144338	11	42	10

As for leukemia, its association with some lncRNAs has been widely reported [28-29]. As illustrated in the following Table 7, it is easy to see that FVTLDA_MLR, FVTLDA_ANN , and KATZLDA can successfully predict 8, 8 and 8 confirmed lncRNAs out of the top 10 candidates lncRNAs separately, and their corresponding case study contrast scores are 0.9448, 0.9753 and 0.9688 respectively.

Table 7: Top 10 potential leukemia-related lncRNAs and their PubMed unique identifiers predicted by FVTLDA_MLR, FVTLDA_ANN and KATZLDA.

disease	lncRNA	Evidence(PMID)	Rank by FVTLDA_MLR	Rank by FVTLDA_ANN	Rank by KATZLDA
leukemia	H19	15645136	1	1	1
leukemia	MALAT1	28713913	2	4	3
leukemia	HOTAIR	26622861	3	2	2
leukemia	PVT1	29510227	4	3	4
leukemia	GAS5	27951730	5	5	5
leukemia	NEAT1	27446393	6	10	8
leukemia	FENDRR	unknown	7	13	14
leukemia	UHRF1	unknown	8	59	69
leukemia	TUG1	29654398	9	6	6
leukemia	XIST	7981672	10	7	9
leukemia	KCNQ1OT1	unknown	18	8	17
leukemia	CCAT1	unknown	14	9	10
leukemia	MIAT	unknown	12	12	7

Moreover, lung cancer is also a leading cause of cancer death all over the world, regardless of gender [30]. As illustrated in the following Table 8, it is easy to see that FVTLDA_MLR and FVTLDA_ANN can successfully predict 8 and 7 confirmed lncRNAs out of the top 10 candidate lncRNAs respectively. However, KATZLDA can only predict 1 confirmed lncRNAs out of the top 10 candidate lncRNAs. Additionally, the case study contrast scores of FVTLDA_MLR, FVTLDA_ANN and KATZLDA are 0.8670, 0.7414 and 0.0998 respectively.

All in all, from the above descriptions, it is obvious that FVTLDA with MLR and FVTLDA with ANN can achieve better prediction performances than KATZLDA, and meanwhile, their corresponding average case study contrast scores are 0.8429 and 0.8515, which are both higher than the average case study contrast score of 0.6375 achieved by KATZLDA as well.

Table 8: Top 10 potential lung cancer-related lncRNAs and their PubMed unique identifiers predicted by FVTLDA_MLR, FVTLDA_ANN and KATZLDA

disease	lncRNA	Evidence(PMID)	Rank by FVTLDA_MLR	Rank by FVTLDA_ANN	Rank by KATZLDA
lung cancer	PVT1	28731781	1	1	4
lung cancer	TUG1	28069000; 29277771	2	2	45
lung cancer	NEAT1	25818739	3	5	51
lung cancer	HOTTIP	26265284	4	6	49
lung cancer	XIST	unknown	5	3	52
lung cancer	DANCR	29651883	6	11	63
lung cancer	MIAT	29487526	7	10	20
lung cancer	KCNQ1OT1	27222340	8	4	60
lung cancer	MIR155HG	unknown	9	7	58
lung cancer	TP53TG1	unknown	10	8	17
lung cancer	HOXA11-AS	29616096	20	9	54
lung cancer	DLX6-AS1	unknown	61	61	1
lung cancer	LINC00511	unknown	62	49	2
lung cancer	GNAS-AS1	unknown	20	43	3
lung cancer	HCG11	unknown	50	55	5
lung cancer	LINC00342	unknown	58	51	6
lung cancer	MIR17HG	unknown	47	62	7
lung cancer	SBF2-AS1	unknown	33	25	8
lung cancer	SNHG12	unknown	46	50	9
lung cancer	HCP5	unknown	51	22	10

Contrast between FVTLDA_MLR and PVTLDA_ANN

In this article, we have proposed two kinds of prediction models such as FVTLDA with MLR and FVTLDA with ANN, in this section, we will further discuss the differences between them in terms

of LOOCV. According to simulation results, we find that the AUC achieved by FVTLDA with MLR is slightly higher than that obtained by FVTLDA with ANN, which may be due to the imperfect design of nodes in the hidden layer of ANN and other parameters in ANN. In the section of case studies, we find that potential lncRNAs predicted by FVTLDA with MLR are similar to those predicted by FVTLDA with ANN. However, in terms of case study contrast scores, FVTLDA with ANN slightly outperforms FVTLDA with MLR. Besides, the execution time of FVTLDA with ANN is a little longer than that of FVTLDA with MLR. Thus, from the above analysis, it is easy to see that either FVTLDA with MLR or FVTLDA with ANN has its advantages while comparing with each other.

Discussion and conclusions

Accumulating evidence have demonstrated that lncRNAs play important roles in the pathological changes of human diseases, and identification of disease-related lncRNAs can help us better understand the disease mechanisms at the molecular level. However, it is costly and time-consuming to verify lncRNA-disease associations with biological experiments. Thus, it is important and necessary to develop efficient computational models to predict potential lncRNA-disease associations.

Different from state-of-the-art prediction models, in this paper, a novel computational model called FVTLDA is proposed to predict potential lncRNA-disease associations based on direct and indirect biological information. In order to avoid the limitation of the single model prediction technique, we further combine FVTLDA with the multiple linear regression and the artificial neural networks respectively. Moreover, in order to evaluate the prediction performance of FVTLDA, we conduct intensive experiments. Simulation results demonstrate that FVTLDA can achieve better performance than these state-of-the-art prediction models. Additionally, in case studies of gastric cancer, leukemia and lung cancer, simulation results show that the prediction ability and stability of both FVTLDA with MLR and FVTLDA with ANN are better than that of competing methods.

Certainly, despite the satisfying prediction performance of FVTLDA, the current version of FVTLDA has some limitations as well. For example, there are five parameters in FVTLDA, how to select the optimal parameter is to be further studied. Additionally, more useful information sources including the gene-disease associations can be integrated into the feature vectors of lncRNA-disease pairs to further improve the prediction performance of FVTLDA in the future.

Materials

In order to introduce direct and indirect biological information on lncRNA-disease associations into FVTLDA, in this section, we first collected three kinds of known associations including miRNA-disease associations, miRNA-lncRNA associations and lncRNA-disease association from various databases. And then, based on these three kinds of datasets, we constructed three kinds of incidence matrix as follows:

Step1: First, we downloaded the dataset of known miRNA-disease associations and miRNA-lncRNA associations from the databases of HMDD [31] and starBase v2.0 [32] respectively. After having removed the repetitive associations supported by multiple evidences,

and normalized the names of the miRNAs in these two datasets, we finally obtained 4704 unique miRNA-disease associations between 246 miRNAs and 373 diseases (see Additional file 1), and 9086 different miRNA-lncRNA association between 246 miRNAs and 1089 lncRNAs (see Additional file 2). Thereafter, based on these two datasets, we constructed a 246×373 dimensional miRNA-disease association incidence matrix *MD* and a 246×1089 dimensional miRNA-lncRNA association incidence matrix *ML* separately. In *MD*, there is $MD(i,j)=1$, if and only if there exists a known association between the miRNA m_i and the disease d_j, otherwise there is $MD(i,j)=0$. Similarly, in *ML*, there is $ML(i,j)=1$, if and only if there exists a known association between the miRNA m_i and the lncRNA l_j, otherwise there is $ML(i,j)=0$. For convenience, we defined the numbers of miRNAs, diseases and lncRNAs obtained above as N_m, N_{d_MD} and N_{l_ML} respectively. Obviously, there are $N_m=246$, $N_{d\_MD}=373$ and $N_{l\_ML}=1089$.

Step2: Next, we downloaded the dataset of known lncRNA-disease associations from the MNDR v2.0 database [33]. After having removed the duplicate associations with multiple evidence, as illustrated in the Figure.11, we further got rid of these associations with either lncRNAs not belonging to N_{l_ML} or diseases not belonging to N_{d_MD}. Finally, we obtained 407 lncRNA-disease associations between 77 different lncRNAs and 95 different diseases (see Additional file 3). similarly, based on the newly-downloaded dataset, we constructed a 77×95 dimensional lncRNA-disease association incidence matrix *LD*, in which, there is $LD(i,j)=1$, if and only if there exists a known association between the lncRNA l_i and the disease d_j, otherwise there is $LD(i,j)=0$. And for convenience, we define the numbers of lncRNAs and diseases obtained above as N_{l_LD} and N_{d_LD} respectively. Obviously, there are $N_{l\_LD}=77$ and $N_{d\_LD}=95$.

Method

1. Construction of the Gaussian Interaction Profile Kernel Similarity for miRNAs based on miRNA-lncRNA associated information

According to the assumption that similar miRNAs tend to interact with similar lncRNAs [34], the Gaussian interaction profile kernel similarity between the miRNA m_i and the miRNA m_j can be calculated as follows:

$$KM(m_i, m_j) = \exp\left(-\gamma_m \left|\left|IP(m_i) - IP(m_j)\right|\right|^2\right) \qquad (4)$$

$$\gamma_m = \frac{\gamma_m'}{\sum_{k=1}^{N_m} ||IP(m_k)||^2} \qquad (5)$$

Here, $IP(m_i)$ denotes the i^{th} row in the miRNA-lncRNA association incidence matrix *ML*, γ_m denotes the normalized bandwidth based on the new bandwidth parameter γ_m', and in this paper γ_m' will be set to 1 according to previous experiments [35]. In this way, an $N_m \times N_m$ dimensional Gaussian interaction profile kernel similarity matrix *KM* for miRNAs can be established.

2. Construction of the Functional Similarity for miRNAs based on miRNA-disease associated information.

In recent years, disease semantic similarity has been widely utilized to identify potential miRNA-disease associations, and many previous researches have shown the validity of this similarity [36]. In this study, we will calculate the disease semantic

similarity in the same way as in previous studies. For all diseases, we will first download its corresponding Medical Subject Headings (MESH) descriptors from the National Library of Medicine in turn (http://www.nlm.nih.gov/) [37], and then, we will represent a disease d_A as its directed acyclic graph (DAG) such as DAG(d_A)=($D(d_A)$, $E(d_A)$). Here, $D(d_A)$ consists of the disease node d_A itself and all ancestor nodes of d_A, while $E(d_A)$ is composed of all the directed edges from parent nodes to children nodes. In the same way of the previous study [17], for any two disease nodes d and t, we will calculate the contribution of t to the semantic value of d as follows:

$$D_d(t) = \begin{cases} 0 & if\ t \notin DAG(d) \\ 1 & if\ t \in DAG(d)\ and\ t = d \\ \max\{\Delta * D_d(t') | t' \in children\ of\ t\} & if\ t \in DAG(d)\ and\ t \neq d \end{cases} \quad (6)$$

Where Δ denotes the semantic contribution decay factor, and according to the previous study [37], in this paper, Δ will be set to 0.5. Thereafter, we can calculate the semantic value of the disease d through combining all these diseases in its DAG(d) as follows:

$$D(d) = \sum_{t_i \in DAG(d)} D_d(t_i) \quad (7)$$

According to the assumption that two diseases with a larger number of shared nodes in their DAGs may have higher similarity, we can calculate the disease semantic similarity score between a pair of diseases d_i and d_j as follows:

$$DS_{md}(i,j) = \frac{\sum_{t \in (DAG(d_i) \cap DAG(d_j))} (D_{d_i}(t) + D_{d_j}(t))}{D(d_i) + D(d_j)} \quad (8)$$

According to the above formula, it is obvious that an $N_{d\_MD} \times N_{d\_MD}$ dimensional matrix DS_{md} can be established. Meanwhile, after extracting the semantic similarity information of disease in the lncRNA-disease association from the matrix DS_{md}, we can further build an $N_{d\_LD} \times N_{d\_LD}$ dimensional matrix DS_{ld} as well.

Apparently, after obtaining the semantic similarity scores of diseases, we can finally obtain the functional similarity between miRNAs based on the assumption that miRNAs with similar functions are often implicated in similar disease [37] as follows: For any two given miRNAs m_i and m_j, let all diseases known to be related to m_i and m_j be $GDM(m_i)=\{d_{i1},d_{i2},d_{i3}...,d_{ip}\}$ and $GDM(m_j)=\{d_{j1},d_{j2},d_{j3},...,d_{jq}\}$ respectively, then the functional similarity score between m_i and m_j can be calculated according to the following:

$$FM(m_i, m_j) = \frac{\sum_{t=1}^{p} \max\left(DS_{md}\left(d_{it}, GDM(m_j)\right)\right) + \sum_{t=1}^{q} \max(DS_{md}(d_{jt}, GDM(m_i)))}{p+q} \quad (9)$$

Obviously, according to the above equation, an $N_m \times N_m$ dimensional functional similarity matrix FM for miRNAs can be established. In the same way, let all diseases are known to be associated to lncRNAs l_i and l_j as $GDL(l_i)=\{d_{i1},d_{i2},d_{i3}...,d_{ip}\}$ and

$GDL(l_j)=\{d_{j1},d_{j2},d_{j3},\ldots,d_{jq}\}$ separately, then the functional similarity score between l_i and l_j can as well be calculated according to the following equation:

$$FL(l_i,l_j)=\frac{\sum_{t=1}^{p}\max\left(DS_{ld}\left(d_{it},GDL(m_j)\right)\right)+\sum_{t=1}^{q}\max(DS_{ld}(d_{jt},GDL(m_i)))}{p+q} \quad (10)$$

3. Construction of FVTLDA

As illustrated in Figure.12, FVTLDA consists of the following three major steps:

Step a: According to indirect biological information including known miRNA-lncRNA associations and known miRNA-disease associations downloaded above, for each pair of lncRNA and disease, a unique feature vector will be constructed first by adopting the random walk with restart based on the Gaussian interaction profile kernel similarity for miRNAs and functional similarity for miRNAs.

Step b: Next, according to known lncRNA-disease associations downloaded above, for each pair of lncRNA and disease, a unique association probability fractions will be calculated based on the concept of Disease Clique.

Step c: Finally, based on the feature vectors and association probability fractions obtained above, the Multiple Linear Regression (MLR) and the Artificial Neural Network (ANN) will be integrated to infer relationships between feature vectors and corresponding association probability fractions. And then, based on these predicted relationships, for each pair of lncRNA and disease, the potential association between them will be mapped into a probability score. Thereafter, based on these probability scores, we can rank the associations between lncRNAs and diseases conveniently.

3.1 Construction of feature vectors for lncRNA-disease pairs

As shown in Figure.2, for each lncRNA-disease pair, the construction of its feature vector consists of the following three major steps:

Step 1: Based on the following formula (11), construct the miRNA-lncRNA association probability fractions matrix *PL* according to known miRNA-lncRNA associations and the Gaussian interaction profile kernel similarity for miRNAs. And then, for each lncRNA l_i, the column corresponding to l_i in the matrix *PL* will be considered as the feature vector of l_i.

Step 2: Based on the following formula (14), construct miRNA-disease association probability fractions matrix *PD* according to known miRNA-disease associations and the miRNA functional similarity. And then, for each disease d_j, the column corresponding to d_j in the matrix *PD* will be considered as the feature vector of d_j.

Step 3: For each lncRNA-disease pair (l_i,d_j), obtain its feature vector through integrating the feature vector of l_i with the feature vector of d_j according to the following formula (17).

Random Walk is usually adopted to sort the association probabilities of nodes in a network [38], therefore we can implement the random walk with restart on the miRNA-lncRNA association network to obtain the feature vector of lncRNAs as follows: Let any given lncRNA node l_i as the walker, the random walks will start from all known miRNA nodes related to it, and will be moved from the current node to the next node according to the Gaussian interaction profile kernel

similarity for miRNA nodes. During implementing the random walk, supposing that the random walk can be restarted with the probability of r_1 ($0<r_1<1$), then the random walk process can be described by the following formulas:

$$PL_{s+1} = (1 - r_1) * NKM^T * PL_s + r_1 * PL_0 \quad (11)$$

$$NKM(i,j) = \frac{KM(i,j)}{\sum_{k=1}^{N_m} KM(i,k)} \quad (12)$$

$$PL_0(i,j) = \frac{ML(i,j)}{\sum_{k=1}^{N_m} ML(k,j)} \quad (13)$$

Obviously, the random walk process is a kind of iterative process, which will be stopped when the random walk reaches a stable state: Here, considering the requirements of time efficiency and accuracy, the random walk will be considered to be stable if the difference between PL_{s+1} and PL_s is less than 10^{-10}. In this way, for each lncRNA l_i, it is obvious that the feature vector of l_i can be expressed by the association probability fractions of all miRNAs related to l_i, i.e., the feature vectors of l_i be expressed by the i^{th} column in the matrix PL.

Similarly, for each disease d_j, let the random walk be restarted with the probability of r_2 ($0 <r_2 <1$), and its feature vector can as well be obtained according to the following equations:

$$PD_{s+1} = (1 - r_2) * NFM^T * PD_s + r_2 * PD_0 \quad (14)$$

$$NFM(i,j) = \frac{FM(i,j)}{\sum_{k=1}^{N_m} FM(i,k)} \quad (15)$$

$$PD_0(i,j) = \frac{MD(i,j)}{\sum_{k=1}^{N_m} MD(k,j)} \quad (16)$$

Finally, for each lncRNA-disease pair (l_i,d_j), its feature vector can be calculated by combining the feature vectors of both l_i and d_j as follows:

$$FV_{ij} = PL(i) \otimes PD(j) \quad (17)$$

Here, $PL(i)$ and $PD(j)$ represent the i^{th} column of the matrix PL and j^{th} column of the matrix PD respectively. Moreover, for two column vectors $A = (a_1, a_2,...,a_n)^T$ and $B = (b_1,b_2,...,b_n)^T$, $A \otimes B=(a_1 \times b_1,a_2 \times b_2,...,a_n \times b_n)^T$.

In this way, all the feature vector obtained will be independent and there is no collinearity.

3.2 Construction of Association Probability Fractions for LncRNA-disease Pair

The incidence matrix LD obtained from known lncRNA-disease associations can only reflect whether or not lncRNAs have known associations with diseases, but cannot accurately express the degrees of their relationships. Moreover, if one element in LD equals 0, it only means that there is currently no known association between the pair of the corresponding lncRNA and disease nodes, but does not mean that there is absolutely no association existing between them. Thus, the values in the matrix LD need to be further processed. Here, we turn this classification problem into a regression problem. By referring to the definition of the Disease Clique proposed in previous study [39], in this section, for each given disease d_i and lncRNA l_j, we will define the set consisting of all these nonzero elements in the i^{th} row of the matrix DS_{ld} as the Disease Clique of d_i, and the set consisting of all these nonzero elements in the j^{th} row of the matrix FL as the LncRNA

Clique of l_j. Then, as shown in the Figure.13, the lncRNA-disease association incidence matrix *LD* can be revised as follows:

$$OUTPUT(i,j) = \frac{OUT(i,j) - \min(OUT)}{\max(OUT) - \min(OUT)} \tag{18}$$

$$OUT = rate * FOUT + (1 - rate) * DOUT \tag{19}$$

$$FOUT(i,j) = \begin{cases} \sum_{n=1}^{N_{l\_LD}} k_1 * LD(n,j) * FL(i,n) & if\ LD(i,j) \neq 1 \\ \left(\sum_{n=1}^{N_{l\_LD}} k_1 * LD(n,j) * FL(i,n)\right) + 1 - k_1 & if\ LD(i,j) = 1 \end{cases} \tag{20}$$

$$DOUT(i,j) = \begin{cases} \sum_{n=1}^{N_{d\_LD}} k_2 * LD(i,n) * DS_{ld}(n,j) & if\ LD(i,j) \neq 1 \\ \left(\sum_{n=1}^{N_{d\_LD}} k_2 * LD(i,n) * DS_{ld}(n,j)\right) + 1 - k_2 & if\ LD(i,j) = 1 \end{cases} \tag{21}$$

The probability fraction matrix *OUTPUT* obtained from the above formula (18) can not only solve the problem of sparsity existing in the original association incidence matrix *LD*, but also reflect the degree of relationships between lncRNAs and diseases to some extent.

3.3 Construction of FVTLDA with MLR and FVTLDA with ANN

In order to avoid the limitations of single model prediction scheme, for any given pair of lncRNA and disease nodes, in this section, we will present two different methods, such as the Multiple linear regression (MLR) analysis and the Artificial neural network (ANN), to reveal the potential relationship between the feature vector of the lncRNA-disease pair and its association probability fraction.

1. Construction of FVTLDA with MLR

MLR analysis is often used in statistical analysis [40-42], whose purpose is to determine the quantitative relationship between the dependent and independent variables, and the general form of MLR can be expressed as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \pm e \tag{22}$$

Here, Y represents the dependent variable, $\{X_1,X_2,...,X_k\}$ denote the independent variable of Y, β_0 is the constant term, $\{\beta_1,\beta_2,...,\beta_k\}$ are the partial regression coefficients of $\{X_1,X_2,...,X_k\}$ respectively, and e denotes the error value. Based on above formula (22), for each lncRNA-disease pair (l_i,d_j), we can represent the relationship between its association probability fraction $OUTPUT(i,j)$ and its feature vector as follows:

$$OUTPUT(i,j) = \beta_0 * 1 + \beta_1 * FV_{ij}(1) + \beta_2 * FV_{ij}(2) + \cdots + \beta_{N_m} * FV_{ij}(N_m) \tag{23}$$

Moreover, for convenience, we define the regression coefficients as $W=[\beta_0,\beta_1,\beta_2,...,\beta_{Nm}]$, the feature vector of (l_i,d_j) as $x_n=[1,FV_{ij}(1),FV_{ij}(2),...,FV_{ij}(N_m)]$, and the association probability fraction corresponding to(l_i,d_j) as $y_n= OUTPUT(i,j)$. Then, for a given training set $T = \{(x_1,y_1),(x_2,y_2),...,(x_N,y_N)\}$, let $X=(x_1,x_2,...,x_n)^T$ and $Y=(y_1,y_2,...,y_n)^T$, the regression coefficients W can be calculated by the least square method. and the optimal solution W^* can be calculated as follows:

$$W^* = (X^T X)^{-1} X^T Y \tag{24}$$

Finally, based on the above formulas, our prediction model FVTLDA with MLR can be described as the following Algorithm 1:

Algorithm 1: FVTLDA with MLR

Input: Known miRNA-lncRNA associations matrix *ML*, known miRNA-disease associations matrix *MD*, known lncRNA-disease associations matrix *LD*, disease MESH descriptors, parameter r_1, r_2, k_1, k_2, *rate*.

Output: The association probability fractions vector of candidate lncRNA-disease pairs.

Step 1: Generate the miRNA Gaussian interaction profile kernel similarity matrix KM by Eq (4);

Step 2: Generate the disease semantic similarity matrix DS_{md} and DS_{ld} by Eq (8);

Step 3: Generate the miRNA function similarity matrix *FM* and lncRNA function similarity matrix *FL* by Eq (9) and Eq (10) respectively;

The training process of the FVTLDA with MLR is as follows:

Step 4: Generate feature vector FV_{ij} for each pair of lncRNA l_i and disease d_j in the training set by Eq (17);

Step 5: Generate the association probability fractions $OUTPUT(i,j)$ for each pair of lncRNA l_i and disease d_j in the training set by Eq (18);

Step 6: Obtain the optimal regression coefficients W∗ by Eq (24);

After obtaining W*, the testing process of the FVTLDA with MLR is as follows:

Step 7: Generate feature vector FV_{ij} for each pair of lncRNA l_i and disease d_j in the testing set;

Step 8: Calculate the association probability fraction $score_{ij}$ for each pair of lncRNA l_i and disease d_j in the testing set as follows: $score_{ij} = W^* \times FV_{ij}$;

Step 9: Sort all candidate lncRNA-disease pairs by the value of association probability fractions obtained by step 8 in the descending order;

2. Artificial neural network (ANN)

ANN is a simple model often used to simulate the biological structure of the human brain. It is a highly dense network composing of simple elements, which can reflect the essential relationships between dependent variables and independent variables. One of the most important characteristics of ANN is that it can be learned by training samples, which can overcome the limitations of traditional methods. Therefore, in this section, we will further adopt ANN to estimate the relationships between the feature vectors of lncRNA-disease pairs and their association probability fractions. As illustrated in the Figure.14, ANN is a parallel distributed processing system composing of many process components (neurons), which can be divided into three layers such as the Input layer, the Hidden layer and the Output layer. In ANN, each neuron in every layer can receive one or more input signals, and generate an output signal through the activation function as the input signal of the next layer. The most import part of ANN is to determine the weights and biases. In ANN, each link between neurons represents a weight that reflects the influence of the previous neuron on the current neuron, and bias can increase the flexibility of this neuron [43]. In

this section, in a way similar to the previous study [44], we determine the weights and biases of ANN through the following four major steps:

Step1: Take the training samples as the input values, and randomly set the initial values of weights and biases in each layer of ANN.

Step2: Calculate the output of ANN and compare the output with the target value to obtain the value of error.

Step3: Readjust the weights and biases in each layer of ANN according to the value of error obtained above from Step 2.

Step4: Repeat the above procedure until ANN reaches the stop condition.

Remarkably, in this paper, all feature vectors of lncRNA-disease pairs will be randomly divided into the training set, the validation set and the test set in a ratio of 3:1:1. Moreover, the training set will be taken as the input of the Input layer, thereafter, the input of the Hidden layer can be obtained by combining the weights, the output of the Input layer and the biases. Additionally, let I_m^n and O_m^n denote the input value and the output value of the node m in the n^{th} layer of ANN separately, then, the output of the Hidden layer can be calculated according to the following activation function:

$$O_X^2 = \frac{2}{1 + e^{-2*I_X^2}} - 1 \qquad (25)$$

Similarly, the input of the Output layer can be acquired by integrating the weights and the output of the Hidden layer, and the output of the Output layer can be figured out through the following activation function:

$$O_1^3 = I_1^3 \qquad (26)$$

After obtaining the output value of the Output layer of ANN, the mean square error (MSE) can be obtained by comparing it with the target (The corresponding association probability fraction) as follows:

$$E_{total} = \frac{1}{N}\sum_{k=1}^{N}(O_1^3(k) - target(k))^2 \qquad (27)$$

Here, N represents the number of test sets.

Finally, the weight and bias between each pair of neuron connections can be modified repeatedly according to the MSE value until one of the following stop conditions has been satisfied:

(1) Maximum training times (will be set to 100 in this paper)
(2) Minimum MSE (will be set to 0.001 in this paper)
(3) Maximum times of consecutive iterations (In the training process, since the MSE of validation set does not decrease in t consecutive iterations, then we will set the maximum times of consecutive iterations to 15 in this paper)

Finally, based on the above formulas, our prediction model FVTLDA with ANN can be described as the following Algorithm 2:

Algorithm 1: FVTLDA with MLR

Input: Known miRNA-lncRNA associations matrix *ML*, known miRNA-disease associations matrix *MD*, known lncRNA-disease associations matrix *LD*, disease MESH descriptors, parameter r_1, r_2, k_1, k_2, *rate*.

Output: The association probability fraction vector of candidate lncRNA-disease pairs.

Step 1: Generate the miRNA Gaussian interaction profile kernel similarity matrix KM by Eq (4);

Step 2: Generate the disease semantic similarity matrix DS_{md} and DS_{ld} by Eq (8);

Step 3: Generate the miRNA function similarity matrix *FM* and lncRNA function similarity matrix *FL* by Eq (9) and Eq (10) respectively;

The training process of the FVTLDA with ANN is as follows:

Step 4: Generate feature vector FV_{ij} for each pair of lncRNA l_i and disease d_j in the training set by Eq (17);

Step 5: Generate the association probability fraction $OUTPUT(i,j)$ for each pair of lncRNA l_i and disease d_j in the training set by Eq (18);

Step 6: Take the FV_{ij} as input values and $OUTPUT(i,j)$ as the target values to generate the weights and biases of ANN according to the four major steps described above;

After obtaining weights and biases of ANN, the testing process of the FVTLDA with ANN is as follows:

Step 7: Generate feature vector FV_{ij} for each pair of lncRNA l_i and disease d_j in the testing set;

Step 8: Take the FV_{ij} as input values to calculate the association probability fraction $score_{ij}$ for each pair of lncRNA l_i and disease d_j in the testing set;

(Here, scoreij is the output values of the ANN.)

Step 9: Sort all candidate lncRNA-disease pairs by the value of association probability fractions obtained by step 8 in the descending order;

Step 10: Output the sorted candidate lncRNA-disease pairs;

List of abbreviations

Feather vectors is develop to predict LncRNA-Disease Associations (FVTLDA); Leave-One Out Cross Validation (LOOCV); Multiple Linear Regression (MLR); Artificial Neural Network (ANN); Cross Validation (CV); Random Walk with Restart (RWR); True Positive Rate (TPR); False Positive Rate (FPR); Receiver Operating Characteristic (ROC); areas under ROC curve (AUC);

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

All data generated or analyzed during this study are included in this published article [Additional file 1, Additional file 2, Additional file 3]

Competing interests

The authors declare that they have no competing interests

Funding

This research was partly sponsored by the National Natural Science Foundation of China (No.61873221, No. 61672447) and the Natural Science Foundation of Hunan Province (No.2018JJ4058, No.2019JJ70010, No.2017JJ5036). Publication costs were funded by the National Natural Science Foundation of China (No.61873221, No.61672447). The funder of manuscript is Lei Wang (L.W.), whose contribution are stated in the section of Author's Contributions. The funding body has not played any roles in the design of the study and collection, analysis and interpretation of data in writing the manuscript.

Author's contributions

YBX conceived the study. YBX, ZX, and LW developed the method. YBX and ZPC implemented the algorithms. XF analyzed the data. YBX and LW wrote the manuscript.

Acknowledgements

The authors thank all those who have made suggestions for this article

Availability and Requirements

Project name: My bioinformatics project FVTLDA
Project home page: https://github.com/xiaoyubin123/FVTLDA.git
Operating system: Platform independent
Programming language: Matlab
Other requirements: Matlab_R2017b or higher
Any restrictions to use by non-academics: No license required

References

1. Esteller, Manel. Non-coding RNAs in human disease[J]. Nature Reviews Genetics, 2011, 12(12):861-874.

2. Wang K C , Chang H Y . Molecular mechanisms of long noncoding RNAs.[J]. Molecular Cell, 2011, 43(6):904-914.

3. Wapinski O , Chang H Y . Long noncoding RNAs and human disease[J]. Trends in Cell Biology, 2011, 21(6):354-361.

4. Mercer T R , Dinger M E , Mattick J S . Long non-coding RNAs: insights into functions[J]. NATURE REVIEWS GENETICS, 2009, 10(3):155-159.

5. Es L , Lm L , Birren B , , et al. Initial sequencing and analysis of the human genome[J]. Nature, 2001, 409(6822):860.

6. Calin G A, Liu C, Ferracin M, et al. Ultraconserved regions encoding ncRNAs are altered in human leukemias and carcinomas[J]. Cancer cell, 2007, 12(3): 215-229.

7.Johnson R . Long non-coding RNAs in Huntington\"s disease neurodegeneration[J]. Neurobiology of Disease, 2012, 46(2):245-254.

8. Cai Y , Yang Y , Chen X , et al. Circulating ‘lncRNA OTTHUMT00000387022’ from monocytes as a novel biomarker for coronary artery disease[J]. Cardiovascular Research, 2016:cvw022.

9. Li J , Xuan Z , Liu C . Long Non-Coding RNAs and Complex Human Diseases[J]. International Journal of Molecular Sciences, 2013, 14(9):18790-18808.

10. Chen G , Wang Z , Wang D , et al. LncRNADisease: a database for long-non-coding RNA-associated diseases[J]. Nucleic Acids Research, 2013, 41(D1):D983-D986.

11. Bu D , Yu K , Sun S , et al. NONCODE v3.0: integrative annotation of long noncoding RNAs[J]. Nucleic Acids Research, 2012, 40(D1):D210-D215.

12. Amaral P P , Clark M B , Gascoigne D K , et al. lncRNAdb: a reference database for long noncoding RNAs[J]. Nucleic Acids Research, 2011, 39(Database):D146-D151.

13. Dinger M E , Pang K C , Mercer T R , et al. NRED: a database of long noncoding RNA expression[J]. Nucleic Acids Research, 2009, 37(Database):D122-D126.

14. Chen X, Yan C C, Zhang X, et al. Long non-coding RNAs and complex diseases: from experimental results to computational models[J]. Briefings in bioinformatics, 2016, 18(4): 558-576.

15. Jingwen Y , Pengyao P , Lei W , et al. A Novel Probability Model for LncRNA–Disease Association Prediction Based on the Naïve Bayesian Classifier[J]. Genes, 2018, 9(7):345-.

16. Yu J, Xuan Z, Feng X, et al. A novel collaborative filtering model for LncRNA-disease association prediction based on the Naïve Bayesian classifier[J]. BMC bioinformatics, 2019, 20(1): 396.

17. Xuan Z, Li J, Yu J, et al. A probabilistic matrix factorization method for identifying lncRNA-disease associations[J]. Genes, 2019, 10(2): 126.

18. Sun J, Shi H, Wang Z, et al. Inferring novel lncRNA–disease associations based on a random walk model of a lncRNA functional similarity network[J]. Molecular BioSystems, 2014, 10(8): 2074-2081.

19. Zhou M, Wang X, Li J, et al. Prioritizing candidate disease-related long non-coding RNAs by walking on the heterogeneous lncRNA and disease network[J]. Molecular BioSystems, 2015, 11(3): 760-769.

20. Chen X. KATZLDA: KATZ measure for the lncRNA-disease association prediction[J]. Scientific reports, 2015, 5: 16840.

21. Liu M X, Chen X, Chen G, et al. A computational framework to infer human disease-associated long noncoding RNAs[J]. PloS one, 2014, 9(1): e84408.

22. Lu C, Yang M, Luo F, et al. Prediction of lncRNA–disease associations based on inductive matrix completion[J]. Bioinformatics, 2018, 34(19): 3357-3364.

23. Chen, Xing. Predicting lncRNA-disease associations and constructing lncRNA functional similarity network based on the information of miRNA[J]. Scientific Reports, 2015, 5:13186.

24. Yang X, Gao L, Guo X, et al. A network based method for analysis of lncRNA-disease associations and prediction of lncRNAs implicated in diseases[J]. PloS one, 2014, 9(1): e87797.
25. Hartgrink H H , Jansen E P M , Grieken N C T V , et al. Gastric cancer.[J]. Lancet, 2009, 374(9688):477-490.
26. Chen D, Ju H, Lu Y, et al. Long non-coding RNA XIST regulates gastric cancer progression by acting as a molecular sponge of miR-101 to modulate EZH2 expression[J]. Journal of experimental & clinical cancer research, 2016, 35(1): 142.
27. Xia H, Chen Q, Chen Y, et al. The lncRNA MALAT1 is a novel biomarker for gastric cancer metastasis[J]. Oncotarget, 2016, 7(35): 56209.
28. Fernando T R, Rodriguez-Malave N I, Waters E V, et al. LncRNA expression discriminates karyotype and predicts survival in B-lymphoblastic leukemia[J]. Molecular cancer research, 2015, 13(5): 839-851.
29. Wang Y, Wu P, Lin R, et al. LncRNA NALT interaction with NOTCH1 promoted cell proliferation in pediatric T cell acute lymphoblastic leukemia[J]. Scientific reports, 2015, 5: 13749.
30. Hoffman P C, Mauer A M, Vokes E E. Lung cancer[J]. Lancet, 2000, 355(9202):479-485.
31. Li Y, Qiu C, Tu J, et al. HMDD v2. 0: a database for experimentally supported human microRNA and disease associations[J]. Nucleic acids research, 2013, 42(D1): D1070-D1074.
32. Li J H, Liu S, Zhou H, et al. starBase v2. 0: decoding miRNA-ceRNA, miRNA-ncRNA and protein–RNA interaction networks from large-scale CLIP-Seq data[J]. Nucleic acids research, 2013, 42(D1): D92-D97.
33. Cui T, Zhang L, Huang Y, et al. MNDR v2. 0: an updated resource of ncRNA–disease associations in mammals[J]. Nucleic acids research, 2017, 46(D1): D371-D374.
34. van Laarhoven T, Nabuurs S B, Marchiori E. Gaussian interaction profile kernels for predicting drug–target interaction[J]. Bioinformatics, 2011, 27(21): 3036-3043.
35. Chen X, Yan G Y. Novel human lncRNA–disease association inference based on lncRNA expression profiles[J]. Bioinformatics, 2013, 29(20): 2617-2624.
36. Chen X, Xie D, Wang L, et al. BNPMDA: bipartite network projection for MiRNA–disease association prediction[J]. Bioinformatics, 2018, 34(18): 3178-3186.
37. Wang D, Wang J, Lu M, et al. Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases[J]. Bioinformatics, 2010, 26(13): 1644-1650.
38. Niu Y W, Wang G H, Yan G Y, et al. Integrating random walk and binary regression to identify novel miRNA-disease association[J]. BMC bioinformatics, 2019, 20(1): 59.
39. Wang L, Xiao Y, Li J, et al. IIRWR: Internal Inclined Random Walk With Restart for LncRNA-Disease Association Prediction[J]. IEEE Access, 2019, 7: 54034-54041.
40. Kaytez F, Taplamacioglu M C, Cam E, et al. Forecasting electricity consumption: A comparison of regression analysis, neural networks and least squares support vector machines[J]. International Journal of Electrical Power & Energy Systems, 2015, 67: 431-438.
41. Atici U. Prediction of the strength of mineral admixture concrete using multivariable regression analysis and an artificial neural network[J]. Expert Systems with applications, 2011, 38(8): 9609-9618.
42. Bahadir E. Using Neural Network and Logistic Regression Analysis to Predict Prospective Mathematics Teachers' Academic Success upon Entering Graduate Education[J]. Educational Sciences: Theory and Practice, 2016, 16(3): 943-964.
43. Lee Y J. Neural network based approach for predicting learning effect in design students[J]. International Journal of Organizational Innovation (Online), 2010, 2(3): 250.
44. Wang L, Zeng Y, Chen T. Back propagation neural network with adaptive differential evolution algorithm for time series forecasting[J]. Expert Systems with Applications, 2015, 42(2): 855-863.

Figure legends

Additional files