BMC Bioinformatics

METHODOLOGY ARTICLE

Open Access

DeepLPI: a multimodal deep learning method for predicting the interactions between lncRNAs and protein isoforms

Dipan Shaw[1*], Hao Chen[1], Minzhu Xie[2*] and Tao Jiang[1,3*]

*Correspondence: dshaw003@ucr.edu; xieminzhu@hotmail.com; jiang@cs.ucr.edu
[1] Department of Computer Science and Engineering, University of California, Riverside, CA 92521, USA
[2] College of Information Science and Engineering, Hunan Normal University, Changsha, China
[3] Bioinformatics Division, BNRIST/Department of Computer Science and Technology, Tsinghua University, Beijing, China

Abstract

Background: Long non-coding RNAs (lncRNAs) regulate diverse biological processes via interactions with proteins. Since the experimental methods to identify these interactions are expensive and time-consuming, many computational methods have been proposed. Although these computational methods have achieved promising prediction performance, they neglect the fact that a gene may encode multiple protein isoforms and different isoforms of the same gene may interact differently with the same lncRNA.

Results: In this study, we propose a novel method, DeepLPI, for predicting the interactions between lncRNAs and protein isoforms. Our method uses sequence and structure data to extract intrinsic features and expression data to extract topological features. To combine these different data, we adopt a hybrid framework by integrating a multimodal deep learning neural network and a conditional random field. To overcome the lack of known interactions between lncRNAs and protein isoforms, we apply a multiple instance learning (MIL) approach. In our experiment concerning the human lncRNA-protein interactions in the NPInter v3.0 database, DeepLPI improved the prediction performance by 4.7% in term of AUC and 5.9% in term of AUPRC over the state-of-the-art methods. Our further correlation analyses between interactive lncRNAs and protein isoforms also illustrated that their co-expression information helped predict the interactions. Finally, we give some examples where DeepLPI was able to outperform the other methods in predicting mouse lncRNA-protein interactions and novel human lncRNA-protein interactions.

Conclusion: Our results demonstrated that the use of isoforms and MIL contributed significantly to the improvement of performance in predicting lncRNA and protein interactions. We believe that such an approach would find more applications in predicting other functional roles of RNAs and proteins.

Background

Long non-coding RNAs (lncRNAs) are RNA transcripts of more than 200 nucleotides that are not translated to proteins. Previous research [1, 2] has demonstrated that lncRNAs participate energetically in almost the whole process of cells. However, the functions of most lncRNAs are unknown. To understand the function of an lncRNA, it is

necessary to identify what other biological molecules it is able to interact with, especially proteins [3, 4]. By interacting with proteins, lncRNAs could regulate the expression of genes, influence nuclear architecture and modulate the activity of proteins [5]. Therefore, identifying lncRNA-protein interactions is an important approach to understand the potential functions of lncRNAs.

Current methods to identify lncRNA-protein interactions are based on biological experiments and computational models. With the rapid development of molecular biology techniques, large-scale experimental approaches such as PAR-CLIP [6], RNAcompete [7], HITS-CLIP [8], and RIP-Chip [9] have been developed to detect RNA-protein binding and have been used to find lncRNA-protein interactions. However, these experimental approaches are expensive and time-consuming [3]. Based on the known lncRNA-protein interactions, many computational methods have been introduced for mining novel lncRNA-protein interactions. According to Zhang et al. [3], the computational methods could be grouped into two broad categories, machine learning-based methods and network-based methods. The machine learning-based methods build binary classifiers to predict lncRNA-protein pairs as interactive or non-interactive. These methods trained their classifiers using sequence, structure and physicochemical features of lncRNAs and proteins. For example, RPISeq [10] utilized the sequence information of RNAs and proteins to train a random forest classifier and a support vector machine classifier. Bellucci et al. trained catRAPID [11] using the physicochemical properties and secondary structure propensities of 592 protein-RNA pairs to predict novel RNA-protein interactions. Wang et al. [12] built a protein-RNA interaction prediction model using the naive Bayes classifier based only on sequence information. LncPro [13] used Fisher's linear discriminant approach to compute a matrix based on lncRNA and protein sequence information, and used the matrix to score the interactions between an lncRNA-protein pair. Based on the sequence and secondary structural information of RNAs and proteins, RPI-Pred [14] trained a support vector machine. RpiCOOL [15] trained a random forest classifier using sequence motifs and repeat patterns. LPI-BLS [16] used sequence information of known lncRNA-protein pairs to learn multiple BLS (broad learning system) classifiers and integrated the classifiers with a logistical regression model. Recently, IPMiner [17], RPI-SAN [18], RPITER [19] and lncADeep [20] employed deep learning techniques to build lncRNA-protein interaction prediction models based on sequence and/or structural information.

Note that, there are several recently developed methods for predicting general ncRNA-protein interactions based on machine learning [21–24], but they do not consider lncRNAs specifically and are hence less relevant to our work.

The network-based methods integrate heterozygous information associated with lncRNAs and proteins into a network [3] and utilize the topological relationship of lncRNAs and proteins to predict lncRNA-protein interactions. Li et al. proposed LPIHN [25] that integrated an lncRNA-lncRNA similarity network, an lncRNA-protein interaction network and a protein–protein interaction (PPI) network into a heterogeneous network, and used a random walk with restart technique on the heterogeneous network to infer lncRNA-protein interactions. Ge et al. developed a different network approach LPBNI [26] using an lncRNA-protein bipartite network inference method. Based on a heterogenous network similar to LPIHN, Xiao et al. proposed PLPIHS [27] using HeteSim

scores [28] to infer lncRNA-protein interactions, and Hu et al. introduced an eigenvalue transformation-based semi-supervised link prediction method LPI-ETSLP [29]. Zhang et al. designed a linear neighborhood propagation method LPLNP [30]. Zhao et al. utilized both random walk and neighborhood regularized logistic matrix factorization and proposed IRWNRLPI [31]. Deng et al. proposed PLIPCOM [32], which combined diffusion and HeteSim features of heterogeneous lncRNA-protein networks and applied a gradient tree boosting algorithm to predict interactions. More recently, [33] combined multiple similarities and multiple features related to lncRNAs and proteins into a feature projection ensemble learning frame. Zhao et al. proposed a semi-supervised learning method LPI-BNPRA [34]. Shen et al. proposed LPI-KTASLP [35], which used multivariate information about lncRNAs and proteins to conduct a semi-supervised link prediction. Xie et al. [36] constructed a network integrating the information about lncRNA expressions, protein–protein interactions and known lncRNA-protein interactions, and adopted a bipartite network recommendation method to predict lncRNA-protein interactions.

Though a lot of computational methods for predicting lncRNA-protein interactions have been introduced, many challenges still remain. First, in the above studies, the machine-learning based methods only focused on the intrinsic features of lncRNAs and proteins and the network based methods mostly focused on the topological features of associated biological networks of lncRNAs and proteins [3]. An integration of all these features might lead to a better prediction. Second, all methods proposed above neglected the fact that a gene may encode multiple protein isoforms and different isoforms of the same gene may interact differently with the same lncRNA, which could inevitably impact their prediction performance. In this paper, we attempt to address these issues and propose a novel method, named DeepLPI (multimodal **Deep** learning method for predicting LncRNA-Protein Isoform interactions). DeepLPI uses sequence, structure and expression data of lncRNAs and protein isoforms. Instead of using the canonical proteins of each gene, DeepLPI considers all protein isoforms, which could help to detect lncRNA-protein interactions more accurately. DeepLPI extracts intrinsic features such as functional motifs from the sequence and structure data, and obtains network topological features from the expression data. Note that, DeepLPI uses mRNA expression data to extract network topological features instead of PPI data as done in the existing methods because most of the available PPI data do not provide the details about isoforms. Moreover, it is possible to build an isoform-isoform interaction network based on mRNA expression data [37].

DeepLPI consists of two parts. In the first part, we train a deep neural network (DNN) that uses the multimodal deep learning (MDL) [38] technique to extract features from the sequence and structure data of lncRNAs and protein isoforms. The MDL fuses these extracted features and measures the initial interaction scores between lncRNAs and protein isoforms. In the second part, a conditional random field (CRF) is designed to exploit the co-expression relationship among lncRNAs and the co-expression relationship among protein isoforms. The CRF assigns final interaction scores between lncRNAs and protein isoforms based on initial interaction scores while trying to keep highly co-expressed lncRNAs and highly co-expressed protein isoforms attaining similar interaction patterns. To overcome the lack of interaction training labels for lncRNAs and

protein isoforms, we propose an iterative semi-supervised training algorithm based on the multiple instance learning (MIL) framework similar to [39–41]. In MIL, for each positive lncRNA and protein interaction pair (r, p), we initially assign positive interaction labels to all pairs (r, i) for each isoform i of p and negative interaction labels to all other pairs of lncRNAs and protein isoforms. In each iteration, the DNN and CRF update the initial interaction scores using co-expressed lncRNAs and co-expressed isoforms until convergence is reached. In this setting, the isoforms of the a protein/gene can interact differently with the same lncRNA. This flexibility and the integration of both intrinsic and network topological features may potentially lead to a better prediction.

To evaluate the performance of DeepLPI, we first measure its prediction performance using protein (i.e., gene) level interactions with lncRNAs provided in the NPInter v3.0 database. We make sure at least a half of our negative interaction examples contain lncRNAs and proteins that are present in the positive interactions (but do not interact with each other). The rest of the negative interactions contain lncRNAs or proteins that are not present in the positive interactions. This helps overcome the overfitting issue. DeepLPI achieved an average AUC (area under receiver operating characteristic curve) of 0.866 and AUPRC (area under the precision-recall curve) of 0.703 on the human interaction dataset. We also compare our method with both machine-learning based methods and network based methods for predicting lncRNA-protein interactions surveyed above on the same dataset. Based on availability and ease of use, 11 methods were chosen for the comparison. The experimental results demonstrate that our method significantly outperformed the others. We further evaluate the effect of various components of our model (i.e., the so-called ablation study), which essentially indicates the effectiveness of each source of data (isoforms, structures, sequences, and expressions) incorporated and how these data are effectively captured by their corresponding components of the model. We also analyze the divergence of isoform interactions, i.e., how isoforms from the same protein may interact differently with lncRNAs. Finally, we validate our method via a series of tests including the correlation similarity test, prediction of mouse lncRNA-protein interactions using the model trained on human data (since lncRNAs are conserved), and case studies of recently discovered lncRNA-protein interactions in the literature.

Results and Discussion

In this section, we first compare the performance of DeepLPI with some state-of-the-art methods and analyze the effectiveness of our method in terms of each type of data we used and each component of the model. Next, we validate the prediction results of DeepLPI using correlation analyses, a mouse dataset as well as some newly discovered human lncRNA-protein interactions in the literature.

Prediction of lncRNA-protein interactions

We first compare the performance of DeepLPI with both of machine-learning based methods and network based methods. Then, we evaluate the effectiveness of each component of our model (i.e., the ablation study). We also study the divergence of lncRNAs interacting with different isoforms of the same protein/gene, and compare the structural components of lncRNAs and protein isoforms in both interactive and non-interactive

pairs. Finally, we test DeepLPI on some smaller and older lncRNA-protein interaction datasets and observe how its performance could be impacted by the size of training data.

Prediction performance comparison between DeepLPI and the existing methods

Since there is no benchmark data for lncRNA-protein isoform interactions, we could only evaluate the performance of DeepLPI based on the benchmark data of (human) lncRNA-protein interactions downloaded from the NPInter v3.0 [42] database. We compare DeepLPI with some state-of-the-art methods including machine-learning based and network based methods.

The popular machine-learning based methods are catRAPID [11], RPISeq [10], lncPro [13], RPI-Pred [14], rpiCOOL [15], IPMiner [17], RPI-SAN [18], lncADeep [20], RPITER [19] and LPI-BLS [16]. Among these methods, some (lncPro and rpiCOOL) provide stand-alone programs, some (catRAPID, RPISeq and RPI-Pred) provide web-based services, some (IPMiner, lncADeep, RPITER and LPI-BLS) are re-trainable with available source codes, while the others are unavailable. Predicting lncRNA-protein interactions on a large scale using web-based services of catRAPID and RPI-Pred is time-consuming and often fails in the case of long input sequences. The publicly available network based methods are LPIHN [25], LPBNI [26], PLPIHS [43], LPLNP [30], PLIPCOM [32], and SFPEL-LPI [33]. Therefore, we compare our method with seven machine-learning based methods lncPro, rpiCOOL, IPMiner, lncADeep, RPITER, LPI-BLS and RPISeq, and six network based methods LPIHN, LPBNI, PLPIHS, LPLNP, PLIPCOM, and SFPEL-LPI. Default parameters of these methods are used as recommended by their authors.

Table 1 shows the average test results in 10 runs of five-fold cross validations on the NPInter v3.0 human dataset. The AUC values of RPISeq (RF), RPISeq (SVM), lncPro, rpiCOOL, IPMiner, lncADeep, RPITER, LPI-BLS and DeepLPI are 0.708, 0.701, 0.723, 0.721, 0.714, 0.825, 0.827, 0.782 and 0.866, respectively, and their AUPRC values are 0.486, 0.473, 0.588, 0.503, 0.569, 0.646, 0.664, 0.575 and 0.685, respectively. DeepLPI

Table 1 Comparison of prediction performance on lncRNA-protein interactions on the NPInter v3.0 human dataset

Broad category	Methods	AUC	AUPRC
Machine-learning based methods	RPISeq (RF)	0.708	0.486
	RPISeq (SVM)	0.701	0.473
	lncPro	0.723	0.588
	rpiCOOL	0.721	0.503
	IPMiner	0.714	0.569
	lncADeep	0.825	0.646
	RPITER	0.827	0.664
	LPI-BLS	0.782	0.575
	DeepLPI	0.866	0.703
Network based methods	LPIHN	0.776	0.421
	LPBNI	0.786	0.559
	PLPIHS	0.672	0.483
	LPLNP	0.801	0.566
	PLIPCOM	0.821	0.609
	SFPEL-LPI	0.823	0.599

outperformed these machine-learning based methods by 22.3%, 23.5%, 19.7%, 20.1%, 21.2%, 4.9%, 4.7%, and 10.7% in terms of AUC and by 44.6%, 48.6%, 19.6%, 39.8%, 23.6%, 8.8%, 5.9%, and 22.2% in terms of AUPRC, respectively. The AUC values of LPIHN, LPBNI, PLPIHS, LPLNP, PLIPCOM and SFPEL-LPI are 0.776, 0.786, 0.672, 0.801, 0.821 and 0.823, respectively, and the AUPRC values are 0.421, 0.559, 0.483, 0.566, 0.609 and 0.599, respectively. DeepLPI also outperformed these network based methods by 11.6%, 10.2%, 28.9%, 8.1%, 5.5% and 5.2% in terms of AUC and 67.0%, 25.8%, 45.5%, 24.2%, 15.4% and 17.4% in terms of AUPRC scores, respectively. Since these results show that DeepLPI, lncADeep and RPITER performed better than the others, we will only compare these three methods in the following experiments.

To test if our sampling method for generating negative interactions is helpful in reducing overfitting, we repeat the above experiment with all negative interactions sampled randomly and compare DeepLPI with two of the best-performing existing methods, lncADeep and RPITER. The AUC values of DeepLPI, lncADeep and RPITER are 0.923, 0.905 and 0.894, respectively, and their AUPRC values are 0.776, 0.753 and 0.747, respectively. While all AUC and AUPRC values of the three methods have increased significantly, DeepLPI consistently performs better than the other two.

We also evaluate the performance of the methods using the leave-one-out cross-validation (LOOCV) experiment, although it is computationally more expensive. In this experiment, the AUC values of DeepLPI, lncADeep and RPITER are 0.855, 0.801 and 0.811, respectively, and their AUPRC values are 0.694, 0.638 and 0.649, respectively. Compared to those in the five-fold cross-validation experiment, the AUC and AUPRC values of all methods decreased a little, which might be due to variance in the data as discussed in [44].

In order to test if homologous protein sequences might have an impact on the performance of DeepLPI and potentially cause data leak and/or model overfitting, we search for homologous proteins in our benchmark dataset based on EggNog [45]. It turns out that only 5% of the proteins are homologous (to other sequences). We repeat the above five-fold cross-validation experiment for DeepLPI by keeping all interactions involving homologous proteins in the same fold. The AUC and AUPRC values decrease only slightly from 0.866 to 0.861 and from 0.703 to 0.699. This suggests that data leak or model overfitting were unlikely or very limited in our experiment.

Analyzing the effects of model components

In order to assess the contribution of the biological features considered in our model as well as its major computational components, we conduct an ablation study by removing various features/components from the model and evaluate how such a change would affect the performance of the model. More specifically, we test how the model is affected when the MIL learning with protein isoforms is replaced by conventional learning with proteins, when the CRF component along with the expression data are removed, and when the sequence or structure data are removed.

Figure 1 shows that the average AUC of DeepLPI dropped 1.4%, 2.0%, 3.3%, and 5.5% without the structure data, without the CRF component for incorporating expression data, without the MIL learning framework for incorporating isoforms, and without the sequence data, respectively. Without these components or data, the performance in

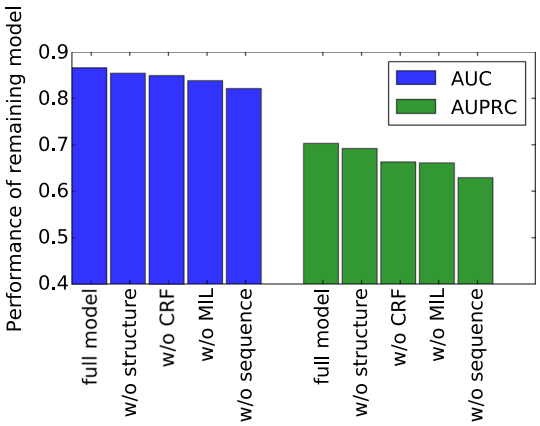

Fig. 1 The effect on performance of removing various components from the model. The average AUC and AUPRC values of DeepLPI, DeepLPI without structure data, DeepLPI without CRF, DeepLPI without using MIL/isoforms, and DeepLPI without sequence data are shown in the figure

term of AUPRC shows a similar declining trend with the percentage decreases being 1.6%, 6.0%, 6.3% and 11.6%, respectively. In particular, when we consider proteins instead of protein isoforms in the model, its AUC dropped from 0.866 to 0.842, which demonstrates the significance of using isoform data. The results also suggest that the CRF component was effective in improving the prediction performance via the integration of expression data. Among all types of input data, the sequences are clearly the most important for the model. Although the usage of structure data did not boost the performance of the model significantly, it allows us to observe interesting enrichment of structural components in interactive lncRNAs, as discussed in the next subsection.

Structural components at important positions in interactive and non-interactive pairs

It would be interesting to study how the structural components of lncRNAs and protein isoforms are distributed in interactive pairs, especially around their interacting sites, and what structural components may contribute more to the interactions than the others. For each lncRNA-protein isoform pair, we use saliency maps [46] to compute importance weights at each position in both sequences. These weights indicate how a position might impact the prediction outcome by our model (i.e., interactive or non-interactive). The lncRNA and protein structural components at heavily weighted (i.e., important) positions of interactive and non-interactive pairs are profiled and shown in Fig. 2a. For each structural component, the average occurrence frequency across all instances is calculated. We can see that at important lncRNA positions, hairpin loops (H) occur much more often in interactive pairs than in non-interactive pairs. The same appears to be true for inner loops (I). On the other hand, stems (S) occur much often at the important lncRNA positions of interactive pairs than at the important lncRNA positions of non-interactive pairs. These suggest that open/unpaired lncRNA positions perhaps play more important roles in their interactions with proteins, and is consistent with several studies in the literature [4, 47].

Similar to lncRNA structural components, we also profile protein isoform structural components in Fig. 2b. However, we are unable to observe a significant difference

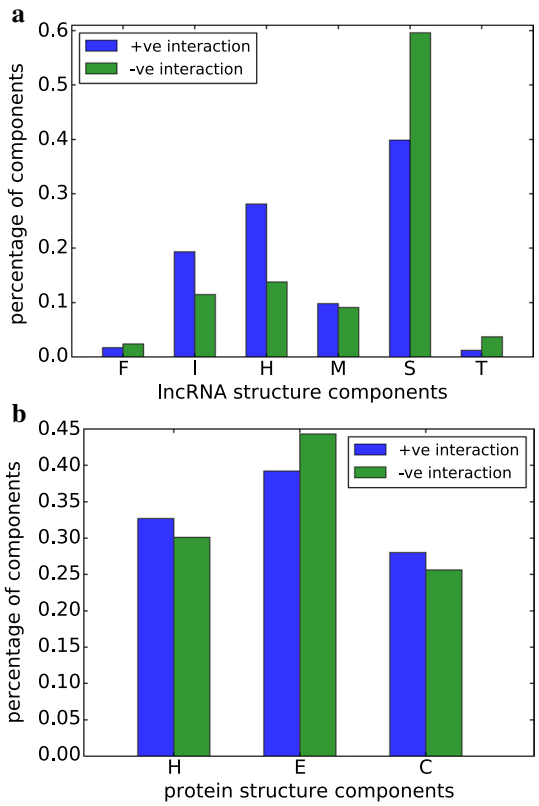

Fig. 2 **a** Distribution of lncRNA structural components. **b** Distribution of protein structural components

between the distributions of the structural components at important protein isoform positions of interactive and non-interactive pairs. We suspect that a more detailed protein structure representation might help reveal some difference, but was unable to pursue it given the time complexity involved in obtaining such representations with high quality.

Divergence of lncRNAs interacting with the isoforms of the same protein

Our ultimate goal is to find lncRNA interactions at the isoform level. Hence, it would be useful to analyze how lncRNAs interact divergently with the isoforms of the same protein. We first estimate the similarity of predicted lncRNA interactions for each pair of isoforms in terms of the semantic similarity score using GOssTo [48]. As in [39, 49], the semantic dissimilarity score between two isoforms is then defined as one minus their similarity score. We consider only proteins with multiple isoforms (MIPs) and collected all interactions between lncRNAs and the isoforms of the MIPs as predicted by DeepLPI trained on the the NPInter v3.0 dataset. For each MIP, the interaction divergence of its isoforms was calculated by averaging the semantic dissimilarity scores of all pairs of its isoforms. Among these MIPs, 71.6% (1548 out of 2163) were estimated to have divergent isoform interactions (i.e., with semantic dissimilarity scores greater than 0). The dissimilarity score distributions for MIPs that have divergent isoform interactions are shown in Additional file 1: Fig. S1 where the mean score value is 0.302.

The impact of training data size

We have collected several (older and smaller) ncRNA-protein interaction datasets including RPI369, RPI1807, RPI2241, and NPInter v2.0. We would like to test how the DeepLPI, lncADeep and RPITER methods perform when these different datasets are used for training and the comparatively newer dataset NPInter v3.0 is used for testing. Since the datasets overlap with each other quite a bit, we make sure that the test interactions do not contain any of the training interactions to prevent a possible data leak. The prediction results are shown in Table 2. The results suggests that the sample size of the training data has a significant effect on the prediction performance of DeepLPI, lncADeep and RPITER. When more training samples are available, these models achieve a better prediction performance, as expected. However, the rate of improvement with respect to the number of interactions is much higher for DeepLPI than for the other two methods.

Validation of predicted lncRNA-protein isoform interactions

To validate the prediction lncRNA-protein isoform interaction results of DeepLPI, we analyze the correlations between isoform sequence similarity, lncRNA sequence similarity, as well as their structure similarity and expression similarity. Moreover, we evaluate the prediction performance of DeepLPI (trained on the NPInter v3.0 human interaction data) using a mouse lncRNA-protein interaction dataset and some new human lncRNA-protein interactions from the recent literature that were not included in the NPInter v3.0 database as the test data.

Correlation analyses

Our basic assumption is that similar lncRNAs tend to interact with similar protein isoforms. To check if our predicted interactions accord to the assumption, we conducted a series analyses of correlation between the similarity of lncRNAs and the similarity of their interactive protein isoforms.

From the lncRNA-protein isoform interactions predicted by DeepLPI, we grouped 1,534 involved lncRNAs into 50 clusters according to a hierarchical clustering based on

Table 2 Performance of DeepLPI, lncADeep and RPITER when datasets from RPI369, RPI1807, RPI2241, and NPInter v2.0 are used for training and the NPInter v3.0 dataset is used for testing

Train \ Test		NPInter v3.0					
Name	#int	DeepLPI		lncADeep		RPITER	
		AUC	AUPRC	AUC	AUPRC	AUC	AUPRC
RPI369	369	0.563	0.202	0.549	0.195	0.551	0.197
RPI1807	1807	0.597	0.271	0.574	0.260	0.580	0.263
RPI2241	2241	0.626	0.328	0.597	0.302	0.609	0.321
NPInter v2.0	6204	0.733	0.513	0.681	0.461	0.693	0.474

Here, #int represents the number of positive lncRNA-protein interactions contained in a training dataset. As the training data increases, the performance DeepLPI, lncADeep and RPITER improves as expected, but the rate of improvement for DeepLPI is higher than the other methods

a generalized Levenshtein (edit) distance. For each group, we calculated a (sequence, structure or expression) similarity score for each pair of lncRNAs in the group and the

average score of the group. We also calculated a similarity score for each pair of protein isoforms that have interaction with the lncRNAs in the group and the average score of all such pairs of protein isoforms. The similarity score between two lncRNA (or protein isoform) sequences is defined as the global alignment score normalized by the alignment length. All similarity scores were normalized to the range of [0, 1]. At last, Pearson's correlation coefficient (PCC) was used to measure the pairwise correlation between lncRNA sequence similarity and protein isoform sequence similarity (Fig. 3a). Similarly, we calculated the PCC between lncRNA expression similarity and protein isoform expression similarity (Fig. 3b), and the PCC between lncRNA structure similarity and protein isoform structure similarity (Fig. 3c). The PCC between lncRNAs sequence similarity and lncRNA expression similarity (Fig. 3d) and the PCC between protein isoform sequence similarity and isoform expression similarity (Fig. 3e) are also included as a useful reference.

Clearly, positive correlations are found in all above analyzes. The strong correlations in Fig. 3a–c conform that similar lncRNAs tend to interact with similar protein isoforms. An interesting observation is that our correlation analysis results are highly consistent with the experimental results in subsection . For example, the strongest correlation between the sequence similarities of lncRNAs and protein isoforms is consistent with the most significant drop in the prediction performance when the sequence data was removed. The moderate correlation coefficients in Fig. 3d, e suggest that the sequence and expression data contain complementary features and thus might explain why their combination helped improve the performance of our model.

Performance on an independent interaction dataset of mouse.

To further validate the effectiveness of DeepLPI in lncRNA-protein interaction prediction, we test DeepLPI and the other existing methods on a dataset independent from the training data. More specifically, we trained all models with the human lncRNA-protein interactions from the NPinter v3.0 database and tested the models on 3580 mouse lncRNA-protein interactions in the same database. Although there is a high genetic similarity between mouse and human (and hence the conservation of lncRNAs), the performance of all models dropped. The AUC of DeepLPI decreased from 0.866 (human) to 0.753 (mouse), but it was still the best since the highest AUC of the other models on the mouse test data was 0.68. An obvious reason for the performance drops might be because lncRNAs do not show the same pattern of evolutionary conservation as protein-coding genes [50].

To further investigate the prediction performance of DeepLPI on interactions between proteins and lncRNAs conserved between human and mouse such as Gas5, Rmst, Neat1 and Meg3 [50, 51], we selected 39 interactions involving conserved lncRNAs from the 3580 mouse interactions. Of the 39 interactions, 89.7% have been correctly predicted by DeepLPI. In particular, since Gas5 is an extensively studied mouse lncRNA that plays an important role in modulating self-renewal [52], we show the interaction prediction results concerning mouse Gas5 in Additional file 1: Table S1. The table demonstrates again that DeepLPI achieved the highest prediction accuracy.

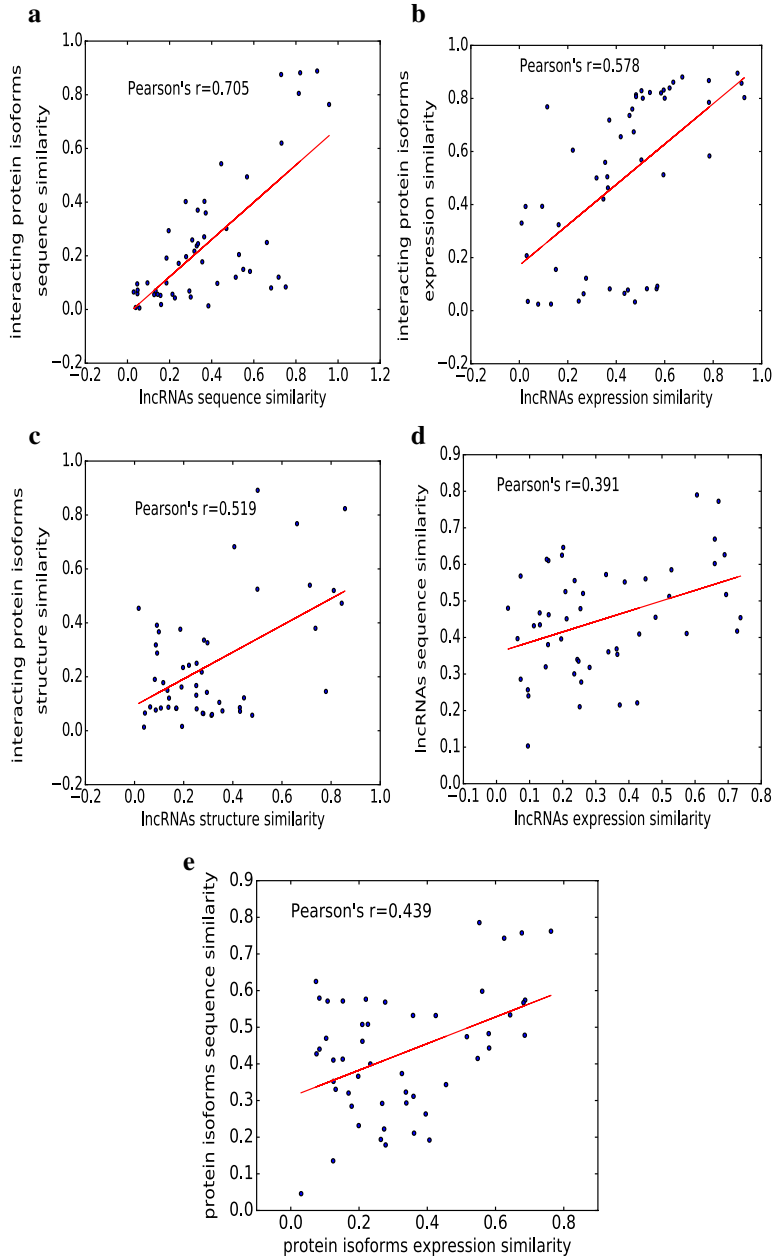

Fig. 3 Correlation analysis. **a** Correlation between lncRNA sequence and protein isoform sequence similarities. **b** Correlation between lncRNA expression and protein isoform expression similarities. **c** Correlation between lncRNAs structure and protein isoform structure similarities. **d** Correlation between lncRNAs sequence and lncRNAs expression similarities. **e** Correlation between protein isoforms sequence and protein isoforms expression similarities

A case study on new interactions

We further validate our model using some new lncRNA-protein interactions from the recent literature that were not included in the NPinter v3.0 database. After a careful literature search, we found 12 new lncRNA-protein interactions [53–56]. The details of these interactions are provided in Additional file 1: Table S2. The prediction results

concerning these new lncRNA-protein interactions by the methods are illustrated in Additional file 1: Table S3. The results show that DeepLPI was able to find out novel interactions often missed by the other methods.

Conclusion

The knowledge of interactions between lncRNAs and protein isoforms could help understand the functions of lncRNAs. In this paper, we proposed a machine-learning based method, DeepLPI, to predict interactions between lncRNAs and protein isoforms. DeepLPI uses a multimodal deep learning neural network to extract intrinsic features from the sequence and structure data of lncRNAs and protein isoforms and a conditional random field to extract network topological features from their expression data. We designed a multiple instance learning iterative algorithm to train the prediction model using an availablelncRNA-protein interaction dataset, and performed extensive experiments to show that DeepLPI achieves a significantly better accuracy in predicting lncRNA-protein interactions compared with the state-of-the-art methods. The multimodal learning feature of DeepLPI allows it to integrate more types of data besides sequences, structures and expression profiles. With minor modifications, DeepLPI could be adapted to predict miRNA-protein interactions, as well as more complex interactions such as lncRNA-miRNA-protein interactions.

Our divergence analysis shows that many isoforms of the same gene interact with different lncRNAs. Hence, it would be of practical importance to study the interactions between lncRNAs and protein isoforms (as opposed to proteins or genes). However, as far as we know, DeepLPI is the first attempt to predict lncRNA-protein isoform interactions, and its performance is still far from being desirable. It might be possible to improve the performance of DeepLPI by using better (e.g., tissue-specific) expression data, more detailed protein secondary structure representations and high quality isoform-isoform interaction network data. We plan to investigate these directions in the near future.

Methods

Datasets

The ground truth interactions between lncRNAs and proteins were downloaded from the NPInter v3.0 database [42]. This is the most enriched database that integrates experimentally verified functional interactions. We kept only the interacting pairs labeled with 'Homo sapiens'. Though the data of NPInter has kept on growing, the number of involved lncRNAs and proteins is still very small at present. In the current version, there are 10031 interactions between 1817 lncRNAs and 151 proteins. These interactions are considered as positive interactions. To train a neural network model, we also need to sample negative interactions that represent pairs of lncRNAs and proteins that do not interact with each other. As the population of negative interactions count is large, complete random sampling of it may contain few lncRNAs and proteins that present in positive interactions, which might lead to overfitting [57]. To reduce overfitting, we make sure that at least a half of the negative lncRNA-protein interaction pairs contain lncRNAs and proteins that appear in positive interaction pairs (but do not interact with each

other). The rest of the negative interaction pairs consist of randomly chosen lncRNAs and proteins that do not appear in positive interaction pairs.

The lncRNA sequences and the protein isoform sequences of human genome were downloaded from GENCODE [58] and ENSEMBL [59], respectively. The sequences were then used to predict their secondary structures. To predict the secondary structure of an lncRNA, we used RNAShapes [60]. The output of RNAShapes was converted to a structure sequence using the EDeN tool (http://github.com/fabriziocosta/EDeN) as in [61]. An lncRNA structure typically consists of six structural components: stem (S), multiloop (M), hairpin loop (H), internal loop (I), dangling end (T), and dangling start (F). To predict the secondary structure from a protein isoform sequence, we used SPIDER2 [62]. SPIDER2 uses a deep neural network to predict a 3-state protein secondary structure whose structural components consist of helix (H), strand (E) and coil (C).

The third type of data that we collected are mRNA and lncRNA expression data. The mRNA expression data are obtained from the literature [49], and the lncRNA expression data were downloaded from the Co-LncRNA database [63]. The mRNA and lncRNA expression data are based on high-throughput human RNA sequencing experiments of 334 studies (1,735 samples) and 241 studies (6,560 samples), respectively. We used the expression data to build co-expression networks. To ensure network quality, we only considered RNA sequencing studies with at least ten samples. Finally, 42 mRNA sequencing studies and 54 lncRNA sequencing studies were kept with a total of 1134 samples and 1429 samples, respectively. Note that an mRNA transcript uniquely corresponds to a protein isoform. In the following, an isoform means either an mRNA transcript or protein isoform. Since different databases use different identifier naming conventions to record protein isoforms, mRNA and lncRNA, ID conversion tools from [63–65] were used to identify the same moleculars from different data sources and perform the mapping between protein isoforms and mRNAs. Finally, we filtered the data and kept the isoforms and lncRNAs that appear in both the sequence data and the expression data.

Data representation

An lncRNA is a character sequence composed of 4 unique ribose nucleotides: cytosine (C), adenine (A), guanine (G), and uracil (U). A protein isoform is a sequence consisting of 20 unique amino acid codes. We generate hexamers and trimers from an lncRNA sequence and a protein isoform sequence, respectively. An lncRNA of length n is represented as $n-5$ consecutive hexamers of ribose nucleotides, and a protein isoform of length n is represented as $n-2$ consecutive trimers of amino acids. A hexamer of nucleotides is encoded as an integer from 0 to 4^6-1, and a trimer of amino acids is encoded as an integer from 0 to 20^3-1. As in [66], to help our deep learning model to learn the intrinsic properties of the sequences efficiently, the integer encoding sequences of lncRNAs and proteins are further encoded using a standard dense embedding technique [67]. A dense embedding maps an integer index of the vocabulary to a dense vector of floats, which is achieved by an embedding layer of our deep learning network using the training data. The embedding layer aims to obtain meaningful dense vectors, which could be utilized to calculate correlations between sequences and are used as the input features of

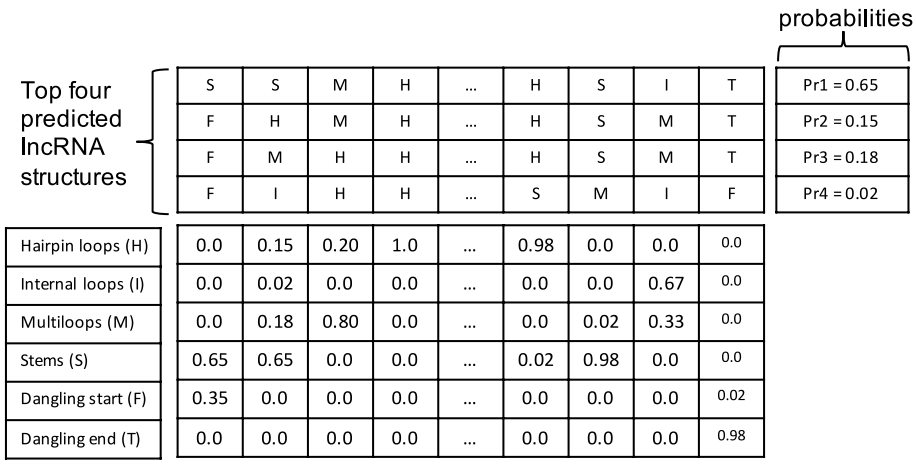

Fig. 4 The representation of multiple predicted structures of an lncRNA. Four predicted structures are merged into a single matrix based on their probabilities

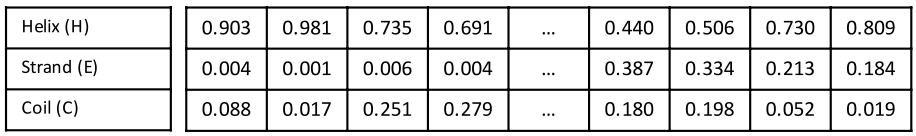

Fig. 5 The representation of multiple predicted structures of a protein. SPIDER2 predicts the probability of each candidate structure, which is summed into a matrix according to structural component types and amino acid positions

lncRNA and protein isoforms. We used a 64-dimensional dense vector to encode a hexamer of nucleotides (or a trimer of amino acids).

Different from the sequence data, the structure of an lncRNA or a protein is often not unique, since multiple structures could be predicted for a single sequence by RNAShapes and SPIDER2. To keep more predicted structural information of an lncRNA of length n, a $6 \times n$ matrix as shown in Fig. 4 is used to encode multiple predicted structures, where the six rows represent six different structural component types and the value at the ith row and jth column is the sum of probabilities of the predicted structures with the jth nucleotide of the lncRNA being of the ith structural component type. Similar to the lncRNA structure representation, a $3 \times n$ probability matrix as shown in Fig. 5 is used to represent multiple predicted structures of SPIDER2 for a protein with n amino acids.

Model architecture and training

DeepLPI predicts the interactions between lncRNAs and protein isoforms by integrating the information of sequence, structure and expression data into a unified predictive model. It consists of two learning submodels. The first is a multimodal deep learning neural network (MDLNN) model and the second a conditional random field (CRF) model. The MDLNN model extracts and fuses the (intrinsic) features from the sequence and structure data of lncRNAs and protein isoforms, and calculates the initial scores of the interactions between lncRNAs and protein isoforms. The CRF model makes a final prediction based on both the initial interaction scores and the expression data of mRNAs

(corresponding to protein isoforms) and lncRNAs. To overcome the lack of ground truth interactions between lncRNAs and proteins, we develop a semi-supervised algorithm following [39–41] to train the MDLNN and CRF models together iteratively. Figure 6 shows a schematic illustration of DeepLPI. More details of the method are described in the following subsections.

Extracting sequence and structure features using multimodal deep learning neural network

To learn intrinsic features related to lncRNA-protein isoform interactions from the sequence and structure data, we construct a multimodal deep learning neural network (MDLNN). We use convolutional layers to extract local features and long short-term memory layer (LSTM) layers to extract short-range and long-range dependencies. At first, MDLNN uses a standard dense embedding technique [67] to map the sequences of lncRNAs and protein isoforms into a 64-dimensional vector space, which is implemented by using embedding layers (denoted as embed(.)) of Keras [68]. After a training

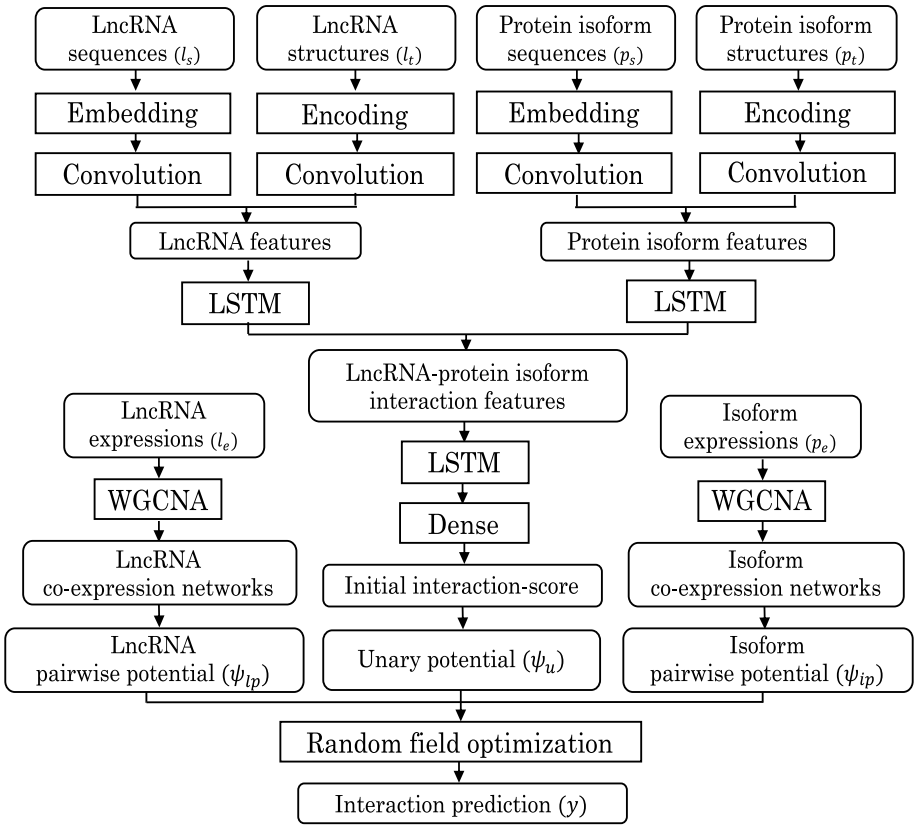

Fig. 6 A flowchart of DeepLPI. It begins with a multimodal deep learning neural network (MDLNN) that uses embedding layers, convolutional layers, LSTM layers and other layers of Keras to extract features from the sequence and structure data of lncRNAs and protein isoforms, and calculate initial interaction scores. Weighted correlation network analysis (WGCNA) is used to construct co-expression networks from expression data of lncRNAs and protein isoforms. Based on the pairwise potentials and unary potentials inferred from the co-expression relationship and the initial interaction scores, respectively, a conditional random field (CRF) optimization is used to predict the interactions between lncRNAs and protein isoforms. The whole model is trained using an iterative semi-supervised learning algorithm based on multiple instance learning (MIL)

process, the embedding layers are able to learn appropriate mappings such that the mapped dense vectors could capture similarities between the sequences. Then, the dense vector matrices representing the sequences and the matrices encoding the predicted structures of lncRNAs and protein isoforms pass through one-dimensional convolutional layers with 4 convolutional filters (denoted as conv(.)) to obtain the local features of the sequence and structure data. After that, max pooling (denoted as pool(.)) layers are used to downsample the output of the convolutional layers to reduce the learning time of the subsequent layers. Based on downsampled features, LSTM layers (denoted as lstm(.)) are used to learn the features that represent the short-range and long-range intrinsic properties of the sequences and structures as in [69, 70]. These features extracted from lncRNA sequences, lncRNA structures, protein isoform sequences and structures are merged together as the input of an LSTM layer followed by a dense layer. The LSTM layer and dense layer (denoted as dense(.)) are intended to learn the interaction patterns between lncRNAs and protein isoforms. Finally, the output of the dense layer is fed into a logistic regression layer (denoted as logit(.)) to compute an initial interaction score. Given an lncRNA sequence l_s, a protein isoform sequence p_s, and the predicted structures l_t and p_t of the lncRNA sequence and protein isoform sequence, respectively, the initial interaction score (IIS) is calculated as follows:

$$\begin{aligned}
&\mathrm{IIS}(l_s, p_s, l_t, p_t)\\
&= \mathrm{logit}(\mathrm{dense}(\mathrm{lstm}(\mathrm{merge}(f_l(f_{l_s}(l_s), f_{l_t}(l_t)), f_p(f_{p_s}(p_s), f_{p_t}(p_t))))))\\
&f_p(f_{p_s}(p_s), f_{p_t}(p_t)) = \mathrm{lstm}(\mathrm{merge}(f_{p_s}(p_s), f_{p_t}(p_t)))\\
&f_l(f_{l_s}(l_s), f_{l_t}(l_t)) = \mathrm{lstm}(\mathrm{merge}(f_{l_s}(l_s), f_{l_t}(l_t)))\\
&f_{l_s}(l_s) = \mathrm{pool}(\mathrm{conv}(\mathrm{embed}(l_s)))\\
&f_{p_s}(p_s) = \mathrm{pool}(\mathrm{conv}((p_s)))\\
&f_{l_t}(l_t) = \mathrm{pool}(\mathrm{conv}(\mathrm{embed}(l_t)))\\
&f_{p_t}(p_t) = \mathrm{pool}(\mathrm{conv}((p_t)))
\end{aligned} \tag{1}$$

Incorporating co-expression relationships using a CRF

Based on the experimental evidence that we have found in the literature, co-expressed isoforms and co-expressed lncRNAs often exhibit similar interactions [71]. To incorporate the co-expression relationships between the isoforms and between the lncRNAs, we use a weighted correlation network analysis (WGCNA) method [72] to construct a co-expression network for the isoforms and one for the lncRNAs separately. In the lncRNA (or protein isoform) co-expression network, the vertices are the lncRNAs (or isoforms, respectively). The edge between vertices i and j has weight $w_{ij} = s_{ij}^{\beta}$, where s_{ij} is the absolute value of the Pearson correlation coefficient (PCC) between the expression profiles of the corresponding lncRNAs (or isoforms) and β is the soft thresholding parameter ($\beta = 6$ in our experiments as suggested by [73] for unsiged networks). Based on the pairwise potentials inferred from the co-expression relationships and the unary potentials inferred from the initial interaction scores output by the MDLNN, DeepLPI next uses a conditional random field (CRF) optimization to predict the interactions between lncRNAs and protein isoforms. Note that our CRF optiimzation framework is very similar to

the framework introduced in [39] for inferring isoform functions. Since many details are different, we will still include a full description of it below for completeness.

For the ith lncRNA-protein isoform pair, denote the lncRNA sequence as l_{s_i}, the protein isoform sequence as p_{s_i}, the lncRNA structure as l_{t_i}, the protein isoform structure as p_{t_i}, the lncRNA expression profile as l_{e_i}, the protein isoform expression profile as p_{e_i}, and the binary label indicating whether there is an interaction between the lncRNA and the protein isoform as y_i. The CRF optimization model aims to obtain the labels y for each lncRNA-protein isoform pair by minimizing the Gibbs energy function below:

$$\begin{aligned} E(y|l_s,p_s,l_t,p_t,l_e,p_e) &= \theta_1 \sum_i \psi_u(y_i|l_{s_i},p_{s_i},l_{t_i},p_{t_i}) + \\ &\theta_2 \sum_{i<j} \psi_{ip}(y_i,y_j|p_{e_i},p_{e_j}) + \theta_3 \sum_{i<j} \psi_{lp}(y_i,y_j|l_{e_i},l_{e_j}) \end{aligned} \tag{2}$$

Here, the Gibbs energy is a weighted summation of unary potentials ψ_u, isoform pairwise potentials ψ_{ip} and lncRNA pairwise potentials ψ_{lp}. The unary potentials ψ_u are calculated from the the initial interaction scores as $\psi_u(0|l_{s_i},p_{s_i},l_{t_i},p_{t_i}) = \mathrm{IIS}(l_s,p_s,l_t,p_t)$ and $\psi_u(1|l_{s_i},p_{s_i},l_{t_i},p_{t_i}) = 1 - \mathrm{IIS}(l_s,p_s,l_t,p_t)$. For the ith and jth lncRNA-protein isoform pairs, their pairwise potential is defined as follows:

$$\begin{aligned} \psi_{ip}(y_i,y_j|p_{e_i},p_{e_j}) &= \mu_p(y_i,y_j) \sum_q w_q(p_{e_i},p_{e_j}) \\ \psi_{lp}(y_i,y_j|l_{e_i},l_{e_j}) &= \mu_p(y_i,y_j) \sum_r w_r(l_{e_i},l_{e_j}) \end{aligned} \tag{3}$$

where $w_q(p_{e_i},p_{e_j})$ is the weight of the edge between isoforms i and j in the q-th isoform co-expression network and $w_r(p_{e_i},p_{e_j})$ is the weight of the edge between lncRNAs i and j in the r-th lncRNA co-expression network. $\mu_p(y_i,y_j)$ is a label compatibility function whose value is 1 if $y_i \neq y_j$ or 0 otherwise. It is used to penalize highly co-expressed isoforms and highly co-expressed lncRNAs assigned with different interaction labels. The weights θ_1, θ_2 and θ_3 are used to control the relative importance of ψ_u, ψ_{ip} and ψ_{lp} in the Gibbs energy. They will be discussed in the next subsection.

By searching for an assignment $\hat{y}$ of labels minimizing the Gibbs energy $E(\hat{y}|l_s,p_s,l_t,p_t,l_e,p_e)$, we attempt to find appropriate labels for lncRNA-protein isoform pairs with low unary energies such that highly co-expressed isoforms would have the same interaction patterns with highly co-expressed lncRNAs. Since computing an exact solution to the Gibbs energy minimization problem is challenging, we apply an efficient approximation algorithm called the mean-field approximation as in [74] to obtain an approximate solution, sketched below.

It is easy to see that minimizing the Gibbs energy is equal to maximizing the following probability:

$$P(y|l_s,p_s,l_t,p_t,l_e,p_e) = \frac{1}{Z} exp(-E(y|l_s,p_s,l_t,p_t,l_e,p_e)) \tag{4}$$

where $Z = \sum_y exp(-E(y|l_s,p_s,l_t,p_t,l_e,p_e))$ is a normalization constant. Let $Q(y|l_s,p_s,l_t,p_t,l_e,p_e)$ be the product of independent marginal probabilities, i.e.,

$$Q(y|l_s,p_s,l_t,p_t,l_e,p_e) = \prod_i Q_i(y_i|l_{s_i},p_{s_i},l_{t_i},p_{t_i},l_{e_i},p_{e_i}) \tag{5}$$

Instead of computing the exact distribution of $P(y|l_s,p_s,l_t,p_t,l_e,p_e)$, we use $Q(y|l_s,p_s,l_t,p_t,l_e,p_e)$ with the minimum KL-divergence $\mathbf{D}(Q||P)$ to approximate P, and adopt the following iterative update equation to obtain a Q with the minimum KL-divergence:

$$\begin{aligned} Q_i(y_i|l_{s_i},p_{s_i},l_{t_i},p_{t_i},l_{e_i},p_{e_i}) = \frac{1}{Z_i} exp\{&-\theta_1\psi_u(y_i|l_{s_i},p_{s_i},l_{t_i},p_{t_i}) \\ &- \theta_2 \sum_{i\neq j}\sum_q w_q(p_{e_i},p_{e_j})Q_j(1-y_i|l_{s_j},p_{s_j},l_{t_j},p_{t_j},l_{e_j},p_{e_j}) \\ &- \theta_3 \sum_{i\neq j}\sum_r w_r(l_{e_i},l_{e_j})Q_j(1-y_i|l_{s_j},p_{s_j},l_{t_j},p_{t_j},l_{e_j},p_{e_j})\} \end{aligned} \tag{6}$$

Here, we initialize Q_i with the unary potential and update it iteratively according to Eq. 6 until convergence, when the final output of our model is obtained.

Training the model with the MIL framework

Because the ground truth lncRNA-protein isoform interactions are generally unavailable, conventional supervised training algorithms cannot be directly applied to our model. Similar to [39] again, here we adopt a semi-supervised training algorithm under the MIL framework as in [40, 41]. In this MIL framework, for each lncRNA, a protein/gene is treated as a bag, the isoforms of a protein/gene are treated as the instances in the bag, and only the ground truth of the bag (i.e., the true lncRNA-protein interaction label) is assumed. We further require that a positive bag should contain at least one positive instance and a negative bag should contain no positive instances. DeepLPI first initializes all instances of positive bags with positive labels, and the other instances with negative labels. Then, the model parameters are optimized with the initial labels in the following standard supervised learning manner.

Given a batch of training instances $(l_s,p_s,l_t,p_t,l_e,p_e,\hat{y})$, the loss functions in terms of the MDLNN parameters w and in terms of the CRF parameters θ are defined as the following negative log likelihoods, respectively.

$$\begin{aligned} \ell_{\text{MDLNN}}(w : l_s,p_s,l_t,p_t,\hat{y}) = -\sum_i (&\hat{y}_i \log(\text{IIS}(l_{s_i},p_{s_i},l_{t_i},p_{t_i})) \\ &+ (1-\hat{y}_i)\log(1-\text{IIS}(l_{s_i},p_{s_i},l_{t_i},p_{t_i}))) \end{aligned} \tag{7}$$

$$\ell_{\text{CRF}}(\theta : l_s,p_s,l_t,p_t,l_e,p_e,\hat{y}) = -\log P(\hat{y}|l_s,p_s,l_t,p_t,l_e,p_e) + \sum_i \frac{\theta_i^2}{2\sigma^2} \tag{8}$$

In Eq. 8, the parameter σ is used to regularize the importance of the co-expression networks in the model optimization and set as 0.1 in our following experiments. We use the Nadam optimization algorithm to update the MDLNN parameters w so ℓ_{MDLNN} could be minimized. To minimize ℓ_{CRF}, we use the L-BFGS-B algorithm as in [39] to update CRF parameters θ.

We perform inference for every instance in the positive bags after each update of the parameters of the model, using the model with the updated parameters. Here, the label of an instance is updated according to the inference: $\hat{y}_i = \text{argmax}_{y_i} P_i(y_i)$. We also adopt the following constraint: for each positive bag, if all its instances are assigned with negative labels, we force the instance with the largest positive prediction score $P_i(1)$ in the bag as positive. The steps of updating parameters and labels are repeated alternately until convergence.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-020-03914-7.

Additional file 1. Supplementary Materials

Abbreviations

LncRNAs: Long non-coding RNAs; PPI: Protein–protein interaction; DNN: Deep neural network; MDL: Multimodal deep learning; CRF: Conditional random field; MIL: Multiple instance learning; AUC: Area under receiver operating characteristic curve; AUPRC: Area under precision-recall curve; MDLNN: Multimodal deep learning neural network; LSTM: Long short-term memory layer; IIS: Initial interaction score; WGCNA: Weighted correlation network analysis; PCC: Pearson correlation coefficient; LOOCV: Leave-one-out cross-validation; MIPs: Proteins with multiple isoforms.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful and constructive comments.

Authors' contributions

D.S. and T.J. designed the model, algorithm and experiments. D.S. implemented the algorithm and conducted the experiments. D.S., H.C. and M.X. wrote the paper. All authors read and approved the final manuscript.

Funding

This work was supported in part by the National Natural Science Foundation of China [61772197], the National Key Research and Development Program of China [2018YFC0910404] and Beijing Natural Science Foundation [4192044]. The funding bodies played no role in the design of the study, or the collection, analysis and interpretation of data, or in writing the manuscript.

Availability of data and materials

DeepLPI is implemented in Python and freely available to the public on https://github.com/dls03/DeepLPI

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 15 July 2020 Accepted: 30 November 2020
Published online: 18 January 2021

References

1. Wilusz JE, Sunwoo H, Spector DL. Long noncoding RNAs: functional surprises from the RNA world. Genes Dev. 2009;23(13):1494–504. https://doi.org/10.1101/gad.1800909.
2. Palazzo AF, Lee ES. Non-coding RNA: what is functional and what is junk? Front Genet. 2015;6:2.
3. Zhang H, Liang Y, Han S, Peng C, Li Y. Long noncoding RNA and protein interactions: from experimental results to computational models based on network methods. Int J Mol Sci. 2019;20(6):1284.
4. Gawronski AR, Uhl M, Zhang Y, Lin Y-Y, Niknafs YS, Ramnarine VR, Malik R, Feng F, Chinnaiyan AM, Collins CC, et al. MechRNA: prediction of lncRNA mechanisms from RNA-RNA and RNA-protein interactions. Bioinformatics. 2018;34(18):3101–10.
5. Kopp F, Mendell JT. Functional classification and experimental dissection of long noncoding RNAs. Cell. 2018;172(3):393–407. https://doi.org/10.1016/j.cell.2018.01.011.
6. Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, Rothballer A, Ascano M Jr, Jungkamp A-C, Munschauer M, Ulrich A, Wardle GS, Dewell S, Zavolan M, Tuschl T. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by par-clip. Cell. 2010;141(1):129–41. https://doi.org/10.1016/j.cell.2010.03.009.

7. Ray D, Kazan H, Chan ET, Peña Castillo L, Chaudhry S, Talukder S, Blencowe BJ, Morris Q, Hughes TR. Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. Nat Biotechnol. 2009;27(7):667–70. https://doi.org/10.1038/nbt.1550.
8. Licatalosi DD, Mele A, Fak JJ, Ule J, Kayikci M, Chi SW, Clark TA, Schweitzer AC, Blume JE, Wang X, Darnell JC, Darnell RB. Hits-clip yields genome-wide insights into brain alternative RNA processing. Nature. 2008;456(7221):464–9. https://doi.org/10.1038/nature07488.
9. Keene JD, Komisarow JM, Friedersdorf MB. Rip-chip: the isolation and identification of mRNAs, microRNAs and protein components of ribonucleoprotein complexes from cell extracts. Nat Protoc. 2006;1(1):302–7. https://doi.org/10.1038/nprot.2006.47.
10. Muppirala UK, Honavar VG, Dobbs D. Predicting RNA-protein interactions using only sequence information. BMC Bioinform. 2011;12(1):489.
11. Bellucci M, Agostini F, Masin M, Tartaglia GG. Predicting protein associations with long noncoding RNAs. Nat Methods. 2011;8(6):444.
12. Wang Y, Chen X, Liu Z-P, Huang Q, Wang Y, Xu D, Zhang X-S, Chen R, Chen L. De novo prediction of RNA-protein interactions from sequence information. Mol BioSyst. 2013;9(1):133–42.
13. Lu Q, Ren S, Lu M, Zhang Y, Zhu D, Zhang X, Li T. Computational prediction of associations between long non-coding RNAs and proteins. BMC Genom. 2013;14(1):651.
14. Suresh V, Liu L, Adjeroh D, Zhou X. Rpi-pred: predicting ncRNA-protein interaction using sequence and structural information. Nucleic Acids Res. 2015;43(3):1370–9.
15. Akbaripour-Elahabad M, Zahiri J, Rafeh R, Eslami M, Azari M. rpicool: A tool for in silico RNA-protein interaction detection using random forest. J Theor Biol. 2016;402:1–8.
16. Fan X-N, Zhang S-W. Lpi-bls: predicting lncRNA-protein interactions with a broad learning system-based stacked ensemble classifier. Neurocomputing. 2019;370:88–93.
17. Pan X, Fan Y-X, Yan J, Shen H-B. Ipminer: hidden ncRNA-protein interaction sequential pattern mining with stacked autoencoder for accurate computational prediction. BMC Genom. 2016;17(1):582.
18. Yi H-C, You Z-H, Huang D-S, Li X, Jiang T-H, Li L-P. A deep learning framework for robust and accurate prediction of ncRNA-protein interactions using evolutionary information. Mol Ther Nucleic Acids. 2018;11:337–44.
19. Peng C, Han S, Zhang H, Li Y. Rpiter: a hierarchical deep learning framework for ncRNA-protein interaction prediction. Int J Mol Sci. 2019;20(5):1070.
20. Yang C, Yang L, Zhou M, Xie H, Zhang C, Wang MD, Zhu H. Lncadeep: an ab initio lncRNA identification and functional annotation tool based on deep learning. Bioinformatics. 2018;34(22):3825–34.
21. Yi H-C, You Z-H, Wang M-N, Guo Z-H, Wang Y-B, Zhou J-R. Rpi-se: a stacking ensemble learning framework for ncRNA-protein interactions prediction using sequence information. BMC Bioinform. 2020;21(1):1–10.
22. Wang L, Yan X, Liu M-L, Song K-J, Sun X-F, Pan W-W. Prediction of RNA-protein interactions by combining deep convolutional neural network with feature selection ensemble method. J Theor Biol. 2019;461:230–8.
23. Zhan Z-H, Jia L-N, Zhou Y, Li L-P, Yi H-C. Bgfe: a deep learning model for ncRNA-protein interaction predictions based on improved sequence information. Int J Mol Sci. 2019;20(4):978.
24. Cheng S, Zhang L, Tan J, Gong W, Li C, Zhang X. Dm-rpis: predicting ncRNA-protein interactions using stacked ensembling strategy. Comput Biol Chem. 2019;83:107088.
25. Li A, Ge M, Zhang Y, Peng C, Wang M. Predicting long noncoding RNA and protein interactions using heterogeneous network model. BioMed Res Int. 2015;2015.
26. Ge M, Li A, Wang M. A bipartite network-based method for prediction of long non-coding RNA-protein interactions. Genom Proteom Bioinform. 2016;14(1):62–71.
27. Xiao Y, Zhang J, Deng L. Prediction of lncRNA-protein interactions using hetesim scores based on heterogeneous networks. Sci Rep. 2017;7(1):3664.
28. Shi C, Kong X, Huang Y, Philip SY, Wu B. Hetesim: a general framework for relevance measure in heterogeneous networks. IEEE Trans Knowl Data Eng. 2014;26(10):2479–92.
29. Hu H, Zhu C, Ai H, Zhang L, Zhao J, Zhao Q, Liu H. Lpi-etslp: lncRNA-protein interaction prediction using eigenvalue transformation-based semi-supervised link prediction. Mol BioSyst. 2017;13(9):1781–7.
30. Zhang W, Qu Q, Zhang Y, Wang W. The linear neighborhood propagation method for predicting long non-coding RNA-protein interactions. Neurocomputing. 2018;273:526–34.
31. Zhao Q, Zhang Y, Hu H, Ren G, Zhang W, Liu H. Irwnrlpi: integrating random walk and neighborhood regularized logistic matrix factorization for lncRNA-protein interaction prediction. Front Genet. 2018;9:239.
32. Deng L, Wang J, Xiao Y, Wang Z, Liu H. Accurate prediction of protein-lncRNA interactions by diffusion and hetesim features across heterogeneous network. BMC Bioinform. 2018;19(1):370.
33. Zhang W, Yue X, Tang G, Wu W, Huang F, Zhang X. Sfpel-lpi: sequence-based feature projection ensemble learning for predicting lncRNA-protein interactions. PLoS Comput Biol. 2018;14(12):1006616.
34. Zhao Q, Yu H, Ming Z, Hu H, Ren G, Liu H. The bipartite network projection-recommended algorithm for predicting long non-coding RNA-protein interactions. Mol Ther Nucleic Acids. 2018;13:464–71.
35. Shen C, Ding Y, Tang J, Jiang L, Guo F. Lpi-ktaslp: prediction of lncRNA-protein interaction by semi-supervised link learning with multivariate information. IEEE Access. 2019;7:13486–96.
36. Xie G, Wu C, Sun Y, Fan Z, Liu J. Lpi-ibnra: long non-coding RNA-protein interaction prediction based on improved bipartite network recommender algorithm. Front Genet. 2019;10:343.
37. Tseng Y-T, Li W, Chen C-H, Zhang S, Chen JJ, Zhou XJ, Liu C-C. Iiidb: a database for isoform-isoform interactions and isoform network modules. BMC Genom. 2015;16(2):10.
38. Ngiam J, Khosla A, Kim M, Nam J, Lee H, Ng AY. Multimodal deep learning. In: Proceedings of the 28th international conference on machine learning (ICML-11), pp. 689–696 (2011)
39. Chen H, Shaw D, Zeng J, Bu D, Jiang T. Diffuse: predicting isoform functions from sequences and expression profiles via deep learning. Bioinformatics. 2019;35(14):284–94.
40. Andrews S, Hofmann T, Tsochantaridis I. Multiple instance learning with generalized support vector machines. In: AAAI/IAAI, pp. 943–944 (2002)
41. Wang X, Yan Y, Tang P, Bai X, Liu W. Revisiting multiple instance neural networks. Pattern Recogn. 2018;74:15–24.

42. Hao Y, Wu W, Li H, Yuan J, Luo J, Zhao Y, Chen R. Npinter v3. 0: an upgraded database of noncoding RNA-associated interactions. Database 2016 (2016)
43. Yang J, Li A, Ge M, Wang M. Prediction of interactions between lncRNA and protein by using relevance search in a heterogeneous lncRNA-protein network. In: 2015 34th Chinese Control Conference (Ccc), pp. 8540–8544 (2015). IEEE
44. Gronau QF, Wagenmakers E-J. Limitations of bayesian leave-one-out cross-validation for model selection. Comput Brain Behav. 2019;2(1):1–11.
45. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, Mende DR, Letunic I, Rattei T, Jensen LJ, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res. 2019;47(D1):309–14.
46. Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013)
47. Wan Y, Qu K, Zhang QC, Flynn RA, Manor O, Ouyang Z, Zhang J, Spitale RC, Snyder MP, Segal E, et al. Landscape and variation of RNA secondary structure across the human transcriptome. Nature. 2014;505(7485):706.
48. Caniza H, Romero AE, Heron S, Yang H, Devoto A, Frasca M, Mesiti M, Valentini G, Paccanaro A. Gossto: a standalone application and a web tool for calculating semantic similarities on the gene ontology. Bioinformatics. 2014;30(15):2235–6.
49. Shaw D, Chen H, Jiang T. Deepisofun: a deep domain adaptation approach to predict isoform functions. Bioinformatics. 2018;35(15):2535–44.
50. Johnsson P, Lipovich L, Grandér D, Morris KV. Evolutionary conservation of long non-coding RNAs; sequence, structure, function. Biochimica et Biophysica Acta (BBA) Gen Subj. 2014;1840(3):1063–71.
51. Li D, Yang MQ. Identification and characterization of conserved lncRNAs in human and rat brain. BMC Bioinform. 2017;18(14):489.
52. Tu J, Tian G, Cheung H-H, Wei W, Lee T-l. Gas5 is an essential lncRNA regulator for self-renewal and pluripotency of mouse embryonic stem cells and induced pluripotent stem cells. Stem Cell Res Ther. 2018;9(1):71.
53. Pospiech N, Cibis H, Dietrich L, Müller F, Bange T, Hennig S. Identification of novel pandar protein interaction partners involved in splicing regulation. Sci Rep. 2018;8(1):2798.
54. Zhang M, Gu Y, Su M, Zhang S, Chen C, Lv W, Zhang Y. Inferring novel lncRNA associated with ventricular septal defect by dna methylation interaction network. BioRxiv. 2018;459677.
55. Yin X, Huang S, Zhu R, Fan F, Sun C, Hu Y. Identification of long non-coding RNA competing interactions and biological pathways associated with prognosis in pediatric and adolescent cytogenetically normal acute myeloid leukemia. Cancer Cell Int. 2018;18(1):122.
56. Xing Y, Zhao Z, Zhu Y, Zhao L, Zhu A, Piao D. Comprehensive analysis of differential expression profiles of mRNAs and lncRNAs and identification of a 14-lncRNA prognostic signature for patients with colon adenocarcinoma. Oncol Rep. 2018;39(5):2365–75.
57. Buda M, Maki A, Mazurowski MA. A systematic study of the class imbalance problem in convolutional neural networks. Neural Netw. 2018;106:249–59.
58. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al. Gencode: the reference human genome annotation for the encode project. Genome Res. 2012;22(9):1760–74.
59. Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T, et al. The ensembl genome database project. Nucleic Acids Res. 2002;30(1):38–41.
60. Steffen P, Voß B, Rehmsmeier M, Reeder J, Giegerich R. RNAshapes: an integrated RNA analysis package based on abstract shapes. Bioinformatics. 2005;22(4):500–3.
61. Pan X, Rijnbeek P, Yan J, Shen H-B. Prediction of RNA-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks. BMC Genom. 2018;19(1):511.
62. Yang Y, Heffernan R, Paliwal K, Lyons J, Dehzangi A, Sharma A, Wang J, Sattar A, Zhou Y. Spider2: a package to predict secondary structure, accessible surface area, and main-chain torsional angles by deep neural networks. In: Prediction of protein secondary structure, pp. 55–63. Springer, Berlin (2017)
63. Zhao Z, Bai J, Wu A, Wang Y, Zhang J, Wang Z, Li Y, Xu J, Li X. Co-lncRNA: investigating the lncRNA combinatorial effects in go annotations and kegg pathways based on human RNA-seq data. Database. 2015;2015.
64. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, et al. Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2015;44(D1):733–45.
65. Fang S, Zhang L, Guo J, Niu Y, Wu Y, Li H, Zhao L, Li X, Teng X, Sun X, et al. Noncodev5: a comprehensive annotation database for long non-coding RNAs. Nucleic Acids Res. 2017;46(D1):308–14.
66. Kulmanov M, Khan MA, Hoehndorf R. Deepgo: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. Bioinformatics. 2017;34(4):660–8.
67. Bengio Y, Ducharme R, Vincent P, Jauvin C. A neural probabilistic language model. J Mach Learn Res. 2003;3(Feb):1137–55.
68. Chollet F, et al.: Keras. https://keras.io (2015)
69. Quang D, Xie X. Danq: a hybrid convolutional and recurrent deep neural network for quantifying the function of dna sequences. Nucleic Acids Res. 2016;44(11):107.
70. Quang D, Xie X. Factornet: a deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data. Methods. 2019;166:40–7.
71. Ehsani R, Drabløs F. Measures of co-expression for improved function prediction of long non-coding RNAs. BMC Bioinform. 2018;19(1):533.
72. Langfelder P, Horvath S. Wgcna: an r package for weighted correlation network analysis. BMC Bioinform. 2008;9(1):559.
73. Langfelder P, Horvath S. Fast R functions for robust correlations and hierarchical clustering. J Stat Softw. 2012;46(11).
74. Krähenbühl P, Koltun V. Efficient inference in fully connected crfs with gaussian edge potentials. In: Advances in neural information processing systems, 2011; pp. 109–117.

Publisher's Note