


METHODOLOGY ARTICLE

Open Access



Capsule-LPI: a LncRNA–protein interaction predicting tool based on a capsule network

Ying Li¹, Hang Sun¹, Shiyao Feng¹, Qi Zhang¹, Siyu Han^{1,2} and Wei Du^{1*} 

*Correspondence:
weidu@jlu.edu.cn

¹ Key Laboratory of Symbolic Computation and Knowledge Engineering, Ministry of Education, College of Computer Science and Technology, Jilin University, Qianjin Street, 130012 Changchun, China
Full list of author information is available at the end of the article

Abstract

Background: Long noncoding RNAs (lncRNAs) play important roles in multiple biological processes. Identifying LncRNA–protein interactions (LPIs) is key to understanding lncRNA functions. Although some LPIs computational methods have been developed, the LPIs prediction problem remains challenging. How to integrate multimodal features from more perspectives and build deep learning architectures with better recognition performance have always been the focus of research on LPIs.

Results: We present a novel multichannel capsule network framework to integrate multimodal features for LPI prediction, Capsule-LPI. Capsule-LPI integrates four groups of multimodal features, including sequence features, motif information, physicochemical properties and secondary structure features. Capsule-LPI is composed of four feature-learning subnetworks and one capsule subnetwork. Through comprehensive experimental comparisons and evaluations, we demonstrate that both multimodal features and the architecture of the multichannel capsule network can significantly improve the performance of LPI prediction. The experimental results show that Capsule-LPI performs better than the existing state-of-the-art tools. The precision of Capsule-LPI is 87.3%, which represents a 1.7% improvement. The F-value of Capsule-LPI is 92.2%, which represents a 1.4% improvement.

Conclusions: This study provides a novel and feasible LPI prediction tool based on the integration of multimodal features and a capsule network. A webserver (<http://csbg-jlu.site/lpc/predict>) is developed to be convenient for users.

Keywords: Long noncoding RNA, lncRNA–protein interaction, Capsule network

Background

Long noncoding RNAs (lncRNAs) are noncoding RNAs that are greater than 200 nt in length and make up the bulk of transcripts [1]. Currently, accumulating research has discovered that lncRNAs play important roles in multiple biological processes [2–5] and are highly associated with diverse human diseases such as tumours and cancers [6–9]. However, the functions and molecular mechanism of the vast majority of lncRNAs remain unknown.

To understand the functions of lncRNAs, there is a fundamental path to identify proteins that interact with lncRNAs. Most lncRNAs need to bind to one or more proteins to



function [10]. Based on the lncRNA–protein interaction (LPI) results, further insights into the functions and molecular mechanisms of lncRNAs can be inferred with the help of abundant annotation information of the protein. Therefore, it is of profound significance to study LPis. There are several ways to explore LPis, which can be divided into experimental methods [11] and computational methods. Experimental methods are time-consuming and expensive [12], while computational methods are efficient and economical.

There are many computational methods of LPI. For instance, Muppirala et al. developed a computational model called RPISeq [12] in 2011, which applies sequence features of lncRNAs and proteins and contains support vector machine (SVM) and random forests (RF) classifiers. In 2013, Lu et al. proposed a method named LncPro [13], which integrates secondary structure features, hydrogen-bonding propensities, and van der Waals interaction features and chooses matrix computation as the calculation method. Then, Suresh et al. proposed RPI-Pred [14] in 2015, which uses sequence features and structure features to develop a model based on SVM. Later, Akbaripour-Elahabad et al. developed RpiCool [15], which utilizes sequence features and motif features and chooses RF as a classifier. In 2015, Li et al. proposed a novel LPI prediction method LPIHN [16] based on random walks with restart on the heterogeneous network constructed by the lncRNA-lncRNA similarity network, lncRNA–protein interaction network, and protein-protein interaction network. In 2016, a network computational method for LPI prediction on LPBNI [17] was developed by Ge et al. In 2017, Zhang et al. developed LPLNP [18], which integrates interaction profile, expression profile, sequence composition features of lncRNAs and interaction profile, CTD features of proteins, and uses the linear neighbourhood similarity and a label propagation process to predict potential LPI. In the same year, Zhang et al. proposed a sequence-based feature learning method called SFPEL-LPI [19]. SFPEL-LPI uses lncRNA sequences, protein sequences, and known lncRNA–protein interactions to compute three lncRNA-lncRNA similarities and protein-protein similarities and combines them with a feature projection ensemble learning frame. In 2018, Zhao et al. proposed a semisupervised model LPI-BNPRA [20], which integrates the lncRNA similarity matrix, protein similarity matrix, and lncRNA–protein interaction matrix to infer LPI. Then, Zhao et al. developed IRWNRLPI [21] for lncRNA–protein interaction prediction by combining random walk algorithms and neighbourhood regularized logistics, which included the lncRNA similarity matrix, protein similarity matrix, and lncRNA–protein interaction matrix. Hu et al. proposed an ensemble method named HLPI-Ensemble [22] by integrating sequence features and the ensemble strategy based on SVM, RF and eXtreme Gradient Boosting. In 2019, Yi et al. developed LPI-Pred [23], which is inspired by the similarity between natural language and biological sequences. LPI-Pred uses word2vec to obtain RNA2vec and Pro2vec as the word embedding features of lncRNAs and proteins, respectively. RF was selected as a classifier to predict LPI.

In recent years, deep learning models have been used for the prediction of LPI. In 2016, Pan et al. developed the computational method IPMiner [24], which employs sequence features and makes use of deep learning to learn hidden features. Then, three RF models were trained, and stacked ensembling was used to integrate different classifiers to further enhance the prediction performance. In 2018, a comprehensive tool

named LncADeep [25] was proposed by Yang et al. In the LPI part, LncADeep integrates the sequence and structure features used in lncPro and some lncRNA features used for lncRNA identification such as Fickett nucleotide features and features of LCDs to infer LPI based on the deep stacking network. In 2020, LPI-CNNCP [26], a novel convolutional neural network method with a copy-padding trick, was proposed by Zhang et al. Zhang et al. also proposed an ensemble deep learning model: lncIBTP [27], which uses sequence features and ensemble CNN and full connection layers as the architecture of lncIBTP. Wekesa et al. proposed a graph representation learning method called GPLPI [28] based on sequence and structural features for LPI prediction. Meanwhile, Wekesa et al. developed a multifeature fusion-based method named DRPLPI [29], which uses a multihead self-attention long short-term memory encoder-decoder network to extract high-level features and feeds them into Catboost and extra tree classification algorithms for LPI prediction.

For most application fields, with the support of large sample sets, deep learning models have better learning performance than traditional machine learning. Deep learning architectures are good at high-level feature extraction, which allows end-to-end learning to be implemented. The design of the deep learning architecture is very flexible. Many deep learning architectures such as CNN [30], DBN [31], RNN [32], BiLSTM [33], attention network [34], capsule network [35] and graph neural network [36] have been developed. The capsule network is one of the most representative networks. To improve the performance of LPI prediction and to explore the effectiveness of the capsule network for LPI, a capsule network has been applied for LPI prediction, which is first proposed for the image recognition field. In the image recognition process, multiple depth features obtained by the feature extraction subnetworks can be well used by the capsule network to make predictions [35]. Compared with other deep learning architectures, the capsule network is more sensitive to the relationship between features. One more advantage of the capsule network over other deep learning architectures is that there are very few parameters to be trained. In our architecture, the capsule network part has only 36 parameters that need to be trained, which makes training faster and improves the overfitting.

Inspired by the better feature-learning capability of the capsule network, in addition to capturing the panorama of LPI information, multiple features are combined, including sequence features, motif information, physicochemical properties and secondary structure features. Another reason for using multimodal features is that lncRNAs and proteins are complex and have many aspects such as sequence information, structural information, and physical and chemical information. Single-modal features have difficulty fully representing lncRNA and protein information, so integrating multimodal features can theoretically produce better prediction performance. At the same time, the advantages of the flexible design of deep learning architectures also create opportunities for the use of multimodal features. For example, Deng et al. proposed a multimodal deep learning framework named DDIMDL [37] in 2020, which constructs deep neural network (DNN)-based submodels to deal with four features and then adopts a joint DNN framework to combine the submodels to make a prediction. In recent years, a variety of LPI prediction methods [22, 28, 29] have adopted multimodal features and achieved good results.

Therefore, we propose a novel multichannel capsule network framework to integrate these multimodal features for LPI prediction, capsule-LPI. The main contributions of Capsule-LPI include:

1. Multimodal features are designed to capture the full information of LPI, including sequence features, motif information, physicochemical properties and secondary structure features. More information features such as physicochemical properties and motif information are integrated into Capsule-LPI compared to existing LPI prediction tools.
2. To better integrate and learn multimodal features, a deep learning architecture based on multichannel capsule networks is proposed to integrate the multimodal features.
3. Capsule-LPI outperforms state-of-the-art methods for LPI prediction with a precision of 87.3% and an F-value of 92.2%. Capsule-LPI also has the significant advantage that very few network parameters need to be trained in the feature-binding part, which makes Capsule-LPI require much less time for training and prediction than other deep learning-based tools.
4. To maximize the convenience for users, a webserver (<http://csbg-jlu.site/lpc/predict>) has been developed. In addition, the source code and dataset used in this paper are provided at <http://csbg-jlu.site/lpc/download>. The source code usage refers to the “README” file in the source code package.

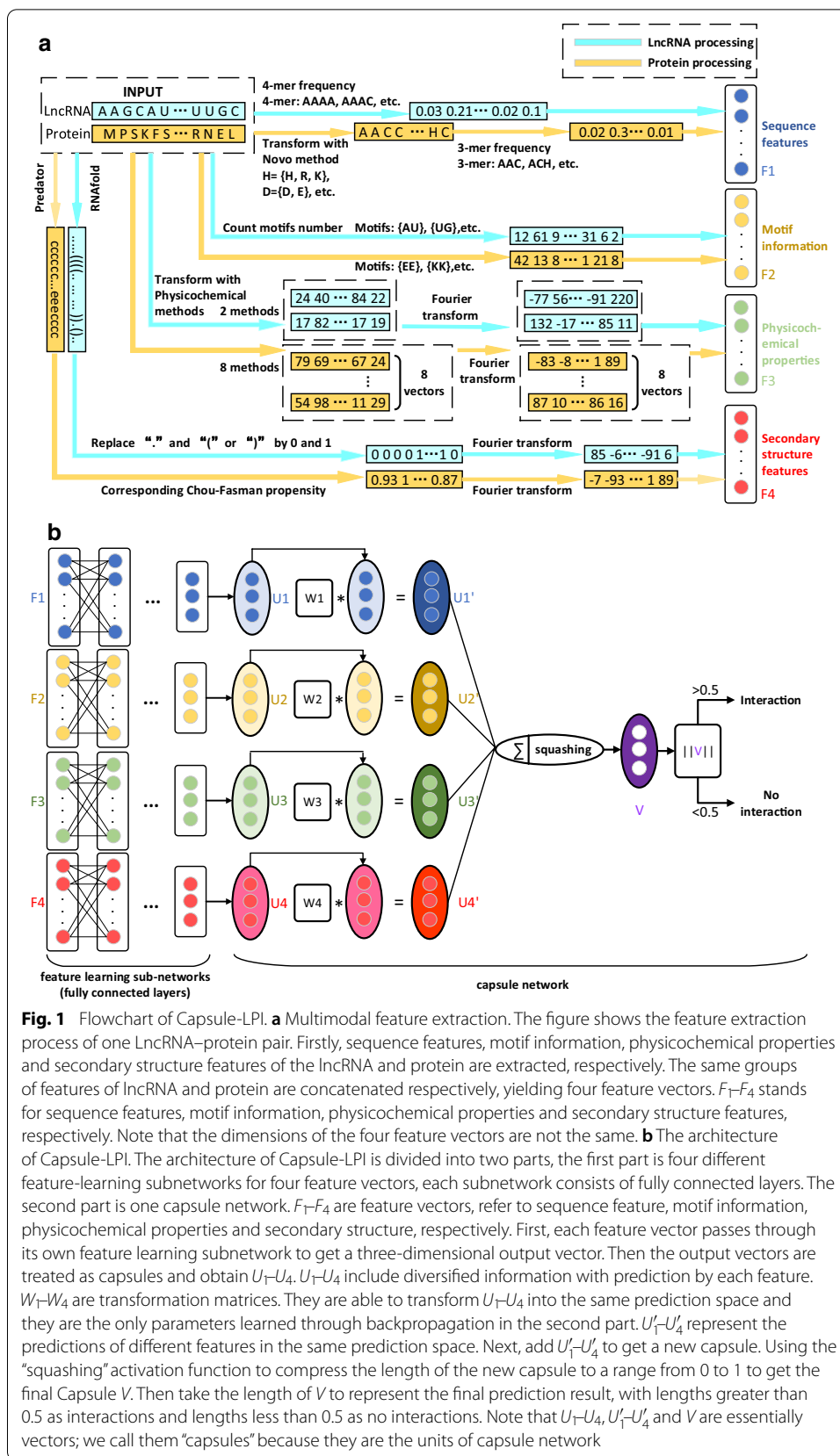
Methods

Capsule-LPI overview

The flowchart of Capsule-LPI is shown in Fig. 1. Capsule-LPI includes two steps: (a) Multimodal feature extraction. When Capsule-LPI received a lncRNA–protein pair, four groups of features, including sequence features, motif information, physicochemical properties and secondary structure features, could be extracted automatically. Each group of features, after being extracted, forms a feature vector. Therefore, four different feature vectors are obtained. (b) The architecture of Capsule-LPI. The architecture of Capsule-LPI consists of four feature-learning subnetworks and one capsule subnetwork [35]. The feature-learning subnetwork consists of fully connected layers. Four different feature vectors obtained by multimodal feature extraction are input into the feature learning subnetworks to automatically learn the more informative and high-level features. Then, capsule subnetworks further integrate the features and predict LPI.

Data description

The dataset is downloaded from NPInter database [38]. The database removes lncRNA–protein interacting pairs for nonhuman species and ncRNAs less than 200 nt in length. A total of 6204 lncRNA–protein interacting pairs were eventually retained. No negative lncRNA–protein samples existed in the database, so we needed to construct negative samples. We use the method of negative sample construction used in existing methods for LPI prediction [15]. The process of generating a negative sample set was as follows: first, all the lncRNAs and proteins used in the positive sample were obtained from the NPInter database, and there were 2356 lncRNAs and 90 proteins in total. Then,



the 2356 lncRNAs and 90 proteins were combined one by one, resulting in a total of 212,040 lncRNA–protein pairs. Finally, the 6204 lncRNA–protein pairs in the positive sample were removed, and 205,836 (21,040–6,204) lncRNA–protein pairs were considered negative samples. The imbalance between large positive samples and negative samples can lead to prediction bias [39], so we randomly divided 205,836 negative samples into 33 sets, and each set contained 6204 negative samples. Thus far, we have obtained 6204 lncRNA–protein interacting pairs and 33 sets consisting of 6204 lncRNA–protein noninteracting pairs, which can be obtained at <http://csbg-jlu.site/lpc/download>. Each lncRNA–protein noninteracting pair set was combined with a lncRNA–protein interacting pair set to train one model. Then, we adopted EasyEnsemble [40] to ensemble 33 models to obtain the final model. More details refer to (Additional file 1: S1).

Multimodal features extraction

To capture more perspectives of LPI, four types of features (sequence features, motif information, physicochemical properties and secondary structure features) are extracted. The extraction process of multimodal features is shown in Fig. 1a. The following four subsections are detailed descriptions of each type of feature.

Sequence features

For lncRNAs, 4-mer frequency features are chosen to encode each lncRNA with a 256 ($4 \times 4 \times 4 \times 4$) dimensional vector, and each element of the vector corresponds to the frequency of the corresponding 4-mer (e.g., AUUC, AACG, CGUC) in the sequence of lncRNAs. The formula for calculating the frequency is as follows:

$$f_i = \frac{n_i}{\sum_{j=1}^{256} n_j} \quad (1)$$

where i is a serial number, f_i is the k -mer frequency of the i -th k -mer and n_i represents the number of i -th k -mer in the sequence.

For proteins, to reduce the feature dimension, the Novo method [22] is used to classify amino acids into four groups: {D, E}, {H, R, K}, {C, G, N, Q, S, T, Y, A}, {F, I, L, M, P, V, W}. Then 3-mer frequency features are chosen to encode each protein with a 64 ($4 \times 4 \times 4$) dimensional vector, and each element of the vector corresponds to the frequency of the corresponding 3-mer in the sequence of the protein. The calculation of frequency refers to Formula 1.

We selected 4-mer frequency features and 3-mer frequency features to encode each lncRNA and protein sequence, respectively, because smaller k -values are poor representations of the sequence, while larger k easily results in sparse representation. In the existing models, the 4-mer frequency features for lncRNA and 3-mer frequency features for protein are mostly considered for LPI prediction [12, 41–43].

In total, the dimension of the sequence feature vector of each lncRNA–protein pair was 320 (256+64).

Motif information

Many motifs have been found to be helpful to predict RNA-protein interactions [44–46]. We use the number of each motif in the sequence to form motif features. Each lncRNA

is encoded with an 18-dimensional vector corresponding to 18 motifs: Fox1, Nova, Slm2, Fusip1, PTB, ARE, hnRNPA1, PUM, U1A, HuD, QKI, U2B, SF1, HuR, YB1, {AU}, {UG} and a motif group, which combines Fox1, Nova, ARE, PUM and U1A. Each protein is encoded with an 11-dimensional vector corresponding to 11 motifs: {H, R}, {HR, RH}, {E}, {K}, {H}, {R}, {EE}, {KK}, {RS, SR}, {RGG} and {YGG}. The details of each motif are provided in (Additional file 1: S2).

In total, the dimension of the motif feature vector of each lncRNA–protein pair was 29 (18 + 11).

Physicochemical properties

The physicochemical properties were used to predict LPI in lncPro [13] and lncADeep [25]. In Capsule-LPI, we adopt the physicochemical properties used in lncPro and add some other physicochemical properties. For lncRNAs, van der Waals interactions and hydrogen-bonding propensities [47] were used to encode each lncRNA sequence into 2 numerical vectors. For proteins, Bull & Breese hydrophobicity [48], Kyte & Doolittle hydrophobicity [49], Zimmerman polarity [50], Grantham polarity [51], isoelectric point, bulkiness, Eisenberg hydrophobicity [52] and Hopp & Woods hydrophobicity propensities [53] were used to encode each protein sequence into 8 numerical vectors. These physicochemical properties are selected because they have been validated by many LPI methods [13, 25].

However, because the dimension of each feature vector depends on the length of the corresponding lncRNA or protein sequence, the input feature vector dimensions of different samples are different. Therefore, the vectors need to be transformed to the same dimension. Here, we adopt the method in lncPro, and use the Fourier transform, which is applied to transform two physicochemical properties into a spectrum domain. The formula of the Fourier series is as follows:

$$X'_k = \sqrt{\frac{2}{L}} \sum_{n=0}^L X_n \cos \left[\frac{\pi}{L} \left(n + \frac{1}{2} \right) \left(k + \frac{1}{2} \right) \right], k = 0, 1, \dots, 9 \quad (2)$$

where L is the length of the original feature vector and X_n is the n -th value in the original feature vector. The first 10 terms of the Fourier series were used as a new spectrum feature vector. Each lncRNA sequence was encoded into two 10-dimensional spectrum vectors corresponding to its two physicochemical feature vectors. Each protein sequence was encoded into eight 10-dimensional spectrum vectors corresponding to its 8 physicochemical feature vectors.

In total, the dimension of the physicochemical spectrum property feature vector of each lncRNA–protein pair was 100 ($2 \times 10 + 8 \times 10$).

Secondary structure features

The secondary structure of lncRNAs and proteins is more conserved than the sequence, which is an important feature to infer LPI. The secondary structure of each lncRNA was obtained using RNAfold [54] based on the minimum free energy algorithm. Then, we transferred the secondary structure to a numerical vector consisting of 0 and 1, in which the paired nucleotide was replaced by 1 and the unpaired nucleotide was replaced by 0.

There is also the problem that the length of the numerical feature vector is related to the length of the sequence, which causes the input vector dimension of different samples to be different. The Fourier transform is used to transform the feature vector and keep the first 10 terms as a new spectrum feature vector. The Fourier series is shown in Formula 2. In this way, the secondary structure feature vector of each lncRNA with dimensions of 10 was obtained.

For the protein secondary structure, the secondary structure sequence of each protein was first obtained using Predator [55]. Then, the secondary structure sequence of each protein was encoded into a numerical feature vector by the Chou-Fasman propensities [56], which were used in the lncPro and lncADeep methods. Each feature vector is also transformed by Fourier transform, and the first 10 terms are retained as a new spectrum feature vector. In this way, the secondary structure feature vector of each protein with dimensions of 10 was obtained.

In total, the dimension of the secondary structure feature vector of each lncRNA–protein pair was 20 (10+10).

Here, the feature vector encoding process was completed. For each lncRNA–protein pair, we obtained 4 groups of feature vectors: sequence feature vector (320-dimensional vector), motif feature vector (29-dimensional vector), physicochemical properties feature vector (100-dimensional vector) and secondary structure feature vector (20-dimensional vector).

Architecture of capsule-LPI

The key architecture of Capsule-LPI is divided into two parts, as shown in Fig. 1b. The first part is four feature-learning subnetworks, and the second part is one capsule subnetwork [35] for prediction. In this section, the architecture of Capsule-LPI and the hyperparameter setting are introduced in detail.

Each feature vector needs one feature-learning subnetwork. Each feature learning subnetwork is made up of fully connected layers, and this subnetwork can not only extract high-level features but also unify the dimensions of feature vectors. The hyperparameters of the feature learning subnetworks are shown in Additional file 1: S3. By experiments, 5 fully connected layers for each subnet are selected because the prediction accuracy does not grow significantly when the layer number is larger than 5, and a larger hidden layer number brings more computation. The number of neurons in each hidden layer was obtained through multiple experiments. PReLU is used as an activation function. To prevent overfitting, we add dropout layers [57] to the hidden layers.

Then, each feature vector is fed into its own feature learning subnet, and the output of each vector is obtained with dimension 3. The dimension 3 was chosen because after trying multiple output dimensions, when the output dimension of the feature extraction subnets is 3, the prediction accuracy of the model is the highest.

The second part of the architecture is a capsule network. The novel learned high-level abstract feature vectors from feature learning subnets are treated as capsules and further fed into the capsule subnetwork. A capsule is essentially a vector; we call it a “capsule” because it is the unit of the capsule network and needs to be distinguished from the vector. A capsule is a group of neurons, and as opposed to a single neuron, a capsule contains more information [35]. As shown in Fig. 1b, U_1-U_4 are capsules corresponding to four high-level

abstract features, which contain multiple predicted information on LPI. W_1-W_4 are transformation matrices that are able to transform U_1-U_4 into the same prediction space and they are the only parameters learned through backpropagation in the second part. $U'_1-U'_4$ are the predictions of different features in the same prediction space. The length of U'_i represents the interaction rate of LncRNA–protein obtained by prediction with the i -th feature, and its direction represents other information on LPI. If $U'_1-U'_4$ are long in length and close in orientation, these properties indicate that the multiple features support LPI in terms of prediction propensity, as well as other interaction information stored in capsules; if $U'_1-U'_4$ are long but differ in orientation, these properties indicate that only the prediction propensity of each feature supports LPI, but other interaction information stored in capsules is not sufficient to support LPI. This network makes the prediction, considering not only the predictive tendencies of each feature but also other interaction information and the relationships between the different features.

To determine whether the capsules ($U'_1-U'_4$) mostly agree with LPI in terms of prediction propensity (reflected in the length of the capsules) as well as other interaction information (reflected in the orientation of the capsules), we add these capsules to obtain a new capsule, S . If S is long, the length of S shows that most of the capsules ($U'_1-U'_4$) are long and the capsules are oriented similarly, indicating that $U'_1-U'_4$ mostly agree with LPI and their other stored information also fits. We do not use the dynamic routing algorithm that is used in the capsule network paper in the adding step because we are a biclassing problem that only needs to output one capsule, which does not require a dynamic routing algorithm. The architecture of the capsule network is shown in (Additional file 1: S3).

Use the length of the final output capsule to represent the LPI's possibility. Therefore, the “squashing” activation function [35] in the capsule network is used to ensure that the short vector shrinks to almost 0 length and the long vector shrinks to a length slightly below 1. The “squashing” activation function formula is:

$$V = \frac{\|S\|^2}{1 + \|S\|^2} \frac{S}{\|S\|} \quad (3)$$

where V is the output capsule, and S is the sum of $U'_1-U'_4$.

Finally, take the length of V to represent the predictions, with lengths greater than 0.5 as interactions and lengths less than 0.5 as no interactions.

Evaluation criteria

To evaluate the performance of Capsule-LPI, we use six evaluation metrics: AUC, AUPRC, accuracy, precision, recall, and F-value. AUC and AUPRC are the area under the ROC and P-R curves, respectively. The formulas for the rest of the evaluation metrics are as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

$$F - value = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (7)$$

where TP , FP , TN and FN represent true positives, false positives, true negatives and false negatives, respectively. TP is the number of samples in the test set for which the prediction result is positive and the label is also positive. FP is the number of samples in the test set for which the prediction result is positive but the label is negative. TN is the number of samples in the test set for which the prediction result is negative and the label is also negative. And FN is the number of samples in the test set for which the prediction result is negative but the label is positive. *Precision* reflects the confidence level when the outcome prediction is positive. *Sensitivity* reflects the probability that we capture the sample when the sample is positive. *Accuracy* and *F - value* are composite measures used to evaluate the comprehensive performance score.

Results

Three experiments have been conducted to evaluate Capsule-LPI in terms of architecture, feature combination, and overall performance.

Architecture comparison

First, it is necessary to evaluate the performance of the architecture of Capsule-LPI. Since Capsule-LPI uses a deep learning architecture, we built three deep learning frameworks, fully connected network (FC), CNN and LSTM, to compare with the architecture of Capsule-LPI. In addition, we also compared the architecture of the existing LPI tool such as the deep stacking network architecture of LncADeep [25]. The architectures were tested on four kinds of features designed in our work (sequence features, motif information, physicochemical properties and secondary structure features). Moreover, a set of control experiments between the architecture of Capsule-LPI and LncADeep using the features of LncADeep was also added to fully assess the performance of the architecture of Capsule-LPI. The features of LncADeep include sequence features and structural features (Additional file 1: S4). The performances of the architecture of Capsule-LPI and other architectures with 10-fold cross-validation are shown in Table 1.

Table 1 shows that under the same features as well as the same test environment, the architecture of Capsule-LPI achieved better performance than the architecture of other deep-learning architectures. On AUC, AUPRC, accuracy, recall and F-value, Capsule-LPI achieves 95.31%, 93.30%, 91.66%, 96.25% and 92.02% under 4 features, respectively, which are all higher than FC, CNN, LSTM and deep stacking network architecture. The architecture of Capsule-LPI has the greatest improvement in the recall metric, which increases close to 3%. The high recall index means that the architecture of Capsule-LPI can identify more potential LPs. The F-value has a nearly 1% increase, indicating that the overall performance of Capsule-LPI is better. To further evaluate whether the improvement of the Capsule-LPI architecture is significant, we calculated the p-values of the F-value between the Capsule-LPI architecture and other deep learning architectures

Table 1 Comparison of the performances of the Capsule-LPI with other deep learning architectures under 10-fold cross-validation

| Tools | AUC (%) | AUPRC (%) | Accuracy (%) | Precision (%) | Recall (%) | F-value (%) |
|---------------------------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|
| FC | 95.19 ± 0.73 | 92.93 ± 1.18 | 90.96 ± 0.86 | 88.75 ± 1.61 | 93.87 ± 0.77 | 91.22 ± 0.76 |
| CNN | 94.64 ± 0.78 | 92.25 ± 1.30 | 89.70 ± 1.22 | 87.24 ± 1.51 | 93.03 ± 1.11 | 90.03 ± 1.14 |
| LSTM | 93.26 ± 0.84 | 88.66 ± 1.64 | 89.70 ± 0.89 | 88.13 ± 1.37 | 91.82 ± 1.94 | 89.91 ± 0.91 |
| LncADeep | 90.55 ± 1.18 | 85.33 ± 2.04 | 87.73 ± 1.03 | 83.89 ± 1.45 | 93.45 ± 1.64 | 88.39 ± 0.96 |
| Capsule-LPI | 95.31 ± 0.41 | 93.30 ± 0.92 | 91.66 ± 0.86 | 88.15 ± 0.86 | 96.25 ± 0.90 | 92.02 ± 0.82 |
| LncADeep (Use LncADeep's features) | 89.52 ± 0.84 | 84.16 ± 1.60 | 87.29 ± 1.04 | 83.70 ± 1.69 | 92.69 ± 1.81 | 87.94 ± 0.97 |
| Capsule-LPI (Use LncADeep's features) | 95.11 ± 0.50 | 93.19 ± 0.65 | 91.34 ± 1.01 | 87.11 ± 1.48 | 97.08 ± 0.85 | 91.82 ± 0.91 |

using the paired t-test on the results of ten iterations. The p-values of F-value for different architecture comparisons are listed as follows: 6.53e-3 for Capsule-LPI vs FC, 1.58e-4 for Capsule-LPI vs CNN, 1.56e-5 for Capsule-LPI vs LSTM, 4.53e-7 for Capsule-LPI vs LncADeep, 8.13e-7 for Capsule-LPI (Use features of LncADeep) vs LncADeep (Use features of LncADeep). All p-values are less than 0.05, which shows that the improvement is significant. The architecture of the performance of Capsule-LPI is also higher than the performance of LncADeep when using the features of LncADeep, indicating that the architecture of Capsule-LPI is not only dependent on the 4 features mentioned in this paper.

Evaluation of combinations of different features

After verifying that the architecture of Capsule-LPI performs well, the multimodal features that are appropriate for Capsule-LPI need to be selected. Here, four features are evaluated. To understand whether each feature is valid in predicting and what combination of features is the best choice, we conducted 15 experiments to evaluate the performance of different features and feature combinations using the Capsule-LPI architecture. The architecture needs to be fine-tuned when inputting different feature combinations. When some features are not adopted, the Capsule-LPI architecture only needs to close the corresponding channel of these features. The architectures of Capsule-LPI for different numbers of feature combinations are shown in Additional file 1: S5. The results for different combinations under 10-fold cross-validation are shown in Table 2.

As shown in Table 2, the single feature of physicochemical properties had the highest recall score, which was 97.83%. The combination of sequence features, motif information, and physicochemical properties yielded the highest AUC score and precision score, which were 95.42% and 88.46%, respectively. For the AUPRC, accuracy and F-value, the combination of 4 features obtained the highest values of 93.30%, 91.66% and 92.02%, respectively. Among the six evaluation indexes, three comprehensive indexes of the combination of 4 features obtained the highest scores, so the combination of 4 features can be considered to be more suitable for the architecture of Capsule-LPI.

Comparison of capsule-LPI performance with existing tools

After verifying the architecture of Capsule-LPI and selecting the suitable feature combinations for Capsule-LPI, the overall performance of Capsule-LPI needs to be evaluated.

Table 2 Comparison of the performance of different features and different feature combinations under 10-fold cross-validation

| Feature Combinations | AUC (%) | AUPRC (%) | Accuracy (%) | Sensitivity (%) | Precision (%) | F-value (%) |
|--------------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|
| SF (Sequence feature) | 94.16 ± 0.79 | 89.99 ± 1.66 | 90.78 ± 0.74 | 87.48 ± 1.20 | 95.20 ± 0.59 | 91.17 ± 0.65 |
| Mtf (Motif information) | 87.97 ± 2.37 | 82.01 ± 4.70 | 85.20 ± 1.28 | 78.99 ± 1.67 | 96.00 ± 0.73 | 86.66 ± 1.04 |
| PC (Physicochemical) | 93.89 ± 0.72 | 90.66 ± 1.47 | 88.99 ± 1.07 | 83.15 ± 1.28 | 97.83 ± 0.97 | 89.89 ± 0.93 |
| SS (Secondary structure) | 89.46 ± 2.38 | 85.35 ± 4.98 | 83.68 ± 1.60 | 78.59 ± 1.79 | 92.67 ± 2.32 | 85.03 ± 1.42 |
| SF + Mtf | 94.43 ± 0.89 | 91.37 ± 1.43 | 91.03 ± 1.03 | 87.33 ± 1.64 | 96.03 ± 0.85 | 91.46 ± 0.91 |
| SF + PC | 95.33 ± 0.63 | 93.42 ± 0.93 | 91.49 ± 1.02 | 87.76 ± 1.71 | 96.48 ± 0.93 | 91.90 ± 0.90 |
| SF + SS | 94.79 ± 0.77 | 92.31 ± 1.41 | 90.83 ± 0.99 | 86.95 ± 1.27 | 96.09 ± 1.00 | 91.29 ± 0.91 |
| Mtf + PC | 94.01 ± 0.80 | 90.82 ± 1.34 | 89.00 ± 1.12 | 83.24 ± 1.60 | 97.74 ± 1.35 | 89.89 ± 0.97 |
| Mtf + SS | 91.37 ± 1.30 | 89.45 ± 1.83 | 84.79 ± 1.10 | 77.96 ± 1.44 | 97.08 ± 0.74 | 86.46 ± 0.83 |
| PC + SS | 94.12 ± 0.61 | 91.66 ± 0.60 | 89.12 ± 1.02 | 83.53 ± 1.19 | 97.50 ± 1.12 | 89.97 ± 0.91 |
| Mtf + PC + SS | 94.13 ± 0.81 | 91.14 ± 1.55 | 89.02 ± 1.18 | 84.14 ± 1.70 | 96.24 ± 2.12 | 89.76 ± 1.09 |
| SF + PC + SS | 95.23 ± 0.65 | 93.17 ± 0.96 | 91.41 ± 1.02 | 86.94 ± 1.43 | 97.48 ± 0.71 | 91.90 ± 0.91 |
| SF + Mtf + SS | 94.73 ± 0.57 | 92.21 ± 0.86 | 90.75 ± 0.82 | 86.98 ± 1.44 | 95.90 ± 0.87 | 91.21 ± 0.71 |
| SF + Mtf + PC | 95.42 ± 0.51 | 93.27 ± 1.00 | 91.62 ± 0.80 | 88.46 ± 1.46 | 95.79 ± 0.98 | 91.96 ± 0.71 |
| SF + Mtf + PC + SS | 95.31 ± 0.41 | 93.30 ± 0.92 | 91.66 ± 0.86 | 88.15 ± 0.86 | 96.25 ± 0.90 | 92.02 ± 0.82 |

Table 3 Comparison of performances for predicting lncRNA–protein interaction by Capsule-LPI and other tools under 5-fold cross validation

| Tools | Sensitivity (%) | Precision (%) | F-value (%) |
|-------------|-------------------|-------------------|-------------------|
| RPISeq(RF) | 99.1 ± 0.2 | 50.1 ± 0.1 | 66.5 ± 0.1 |
| RPISeq(SVM) | 93.5 ± 0.7 | 50.2 ± 0.2 | 65.3 ± 0.4 |
| lncPro | 80.3 ± 0.9 | 52.2 ± 0.6 | 63.2 ± 0.4 |
| RPI-pred | 88.0 ± 0.3 | 49.8 ± 0.6 | 63.6 ± 0.5 |
| rpiCool | 92.0 ± 0.8 | 83.3 ± 0.8 | 87.5 ± 0.6 |
| IPMiner | 89.8 ± 1.1 | 85.6 ± 0.7 | 87.6 ± 0.6 |
| LncADeep | 97.0 ± 0.5 | 85.4 ± 0.8 | 90.8 ± 0.4 |
| Capsule-LPI | 97.6 ± 0.6 | 87.3 ± 0.2 | 92.2 ± 0.3 |

Several state-of-the-art tools for predicting RNA–protein interactions are compared, i.e., RPISeq [12], lncPro [13], RPI-pred [14], rpiCool [15], IPMiner [24] and LncADeep [25]. For the experiments to be comparable, Capsule-LPI is evaluated concerning the methodology in LncADeep, in which the same dataset used in LncADeep, as well as the evaluation method, are adopted. For the same data set, Capsule-LPI uses the exact same positive sample as LncADeep. Since LncADeep does not provide the negative sample, Capsule-LPI uses the same negative sample generation method as LncADeep. For the same evaluation method, 5-fold cross-validation, which is used in LncADeep, is used to evaluate the performance of Capsule-LPI. The comparison results are shown in Table 3.

As shown in Table 3, under the 5-fold cross validation averaging assessment condition, Capsule-LPI with 87.3% precision and 92.2% F-value is superior to other existing

tools. Since other tools do not support retraining, we only calculated the AUC and AUPRC of Capsule-LPI under 5-fold cross validation, which were $95.28 \pm 0.47\%$ and $95.26 \pm 0.62\%$, respectively. The precision of Capsule-LPI was at least 1.7% higher than the precision of the other tools, indicating that the LncRNA–protein interaction pairs obtained by Capsule-LPI prediction are highly reliable. The F-value of Capsule-LPI also achieves a 1.4% improvement, showing that the overall performance of Capsule-LPI is the best. For sensitivity, Capsule-LPI obtains 97.6%, which is slightly lower than the highest sensitivity of 99.1% obtained by RPISeq (RF). However, RPISeq does not perform well on the other two metrics, and its precision is only 50.1%. Therefore, overall Capsule-LPI outperforms the outstanding current tools.

Discussion

In the results section, three experiments have evaluated the performance of Capsule-LPI. First, to verify the architecture of Capsule-LPI, it was tested against four architectures. The experimental results show that the architecture of Capsule-LPI outperforms these four architectures, which shows that it is an effective architecture. Second, a comprehensive feature evaluation experiment is conducted. We consider four kinds of features used in the existing LPI prediction tools and select the best combination of features for Capsule-LPI, which contains sequence features, motif information, physicochemical properties and secondary structure features. Finally, Capsule-LPI is compared to other outstanding LPI prediction tools. The results show that Capsule-LPI outperforms state-of-the-art methods in LPI prediction.

However, a good tool should not only have good performance but also be helpful to scientific research and easy to use. In this regard, we have done a case study to introduce how this work can help with the research of lncRNAs and develop a webserver that is convenient to use.

One case study: finding lncRNA related diseases

To demonstrate the effectiveness and practicability of Capsule-LPI for follow-up research on lncRNAs, a case study for lncRNA-disease association was conducted. In this case, study, we used Capsule-LPI to predict which proteins interact with the top 10 lncRNAs of interest on PubMed and less studied lncRNAs on PubMed. Then, for each lncRNA, the diseases were inferred according to the enriched diseases of the predicted interacting proteins computed by hypergeometric distribution inference. The process of the case study is as follows:

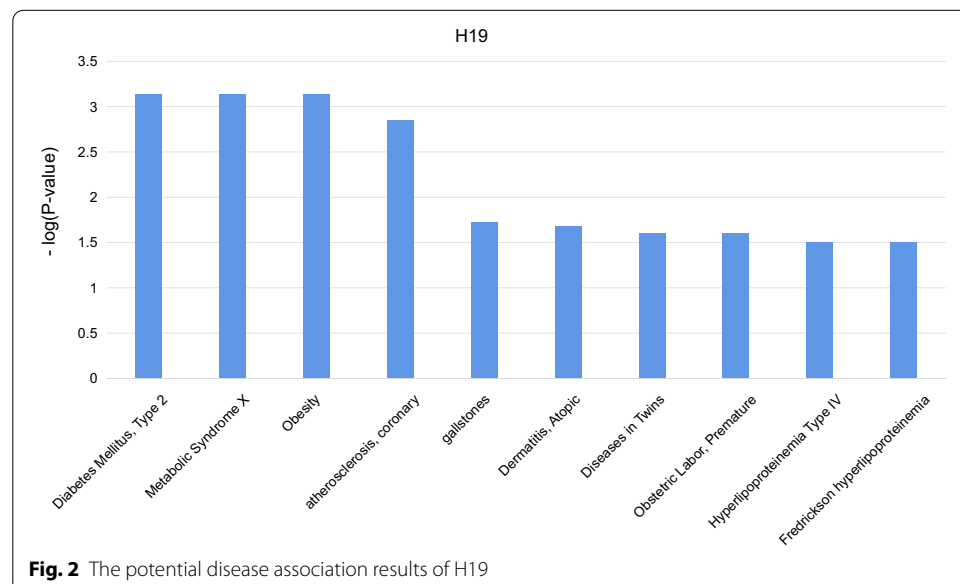
1. We queried the PubMed database to obtain H19, MALAT1, HOTAIR, MEG3, NEAT1, GAS5, UCA1, XIST, PVT1 and TUG1 and downloaded the sequences of these lncRNAs.
2. A total of 26,560 protein sequences of Homo sapiens were downloaded from the UniProt database.
3. Capsule-LPI was used to predict the interacting proteins for lncRNAs, in which the threshold value was set to 0.87 for the higher confidence of the LPI results.
4. The potential disease association of each lncRNA was inferred by the enrichment analysis of disease association for the interacting proteins through the DAVID online

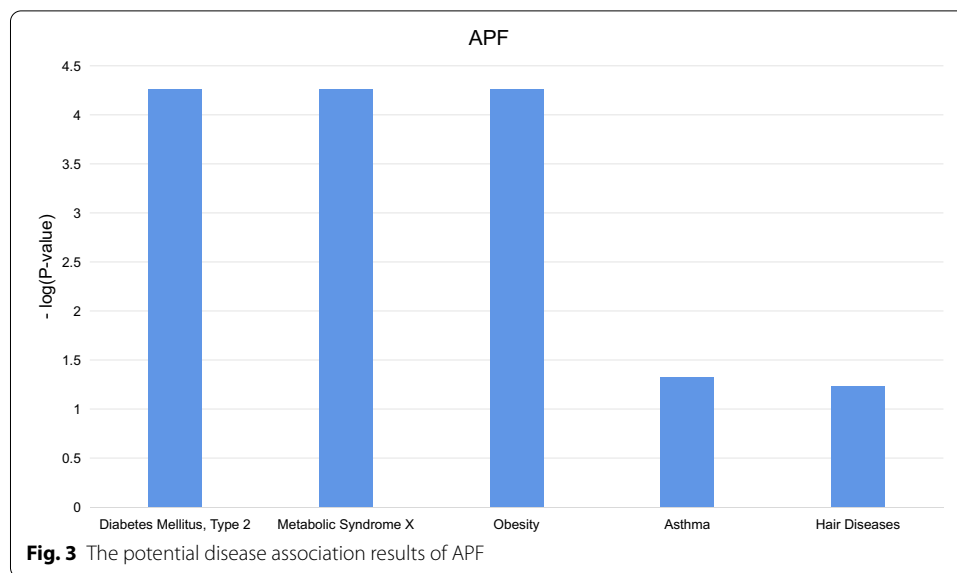
service website. Here, we show the potential disease association results of H19 in Fig. 2, and the rest of the lncRNAs are referred to (Additional file 1: S6).

Furthermore, potentially related diseases are categorized into different types of diseases, including metabolic, chemical dependence, cardiovascular, immune, pharmacogenomic, cancer, infection, neurological, renal, ageing, developmental, haematological, psych, reproduction and vision. The top 10 lncRNAs on PubMed are highly related to metabolic, chemical dependence, cardiovascular, immune and pharmacogenomic processes. To test the validity of this result, we searched the LncRNADisease database for the diseases corresponding to each lncRNA, and the vast majority of the diseases were covered by our predictions. For example, the diseases currently known to be associated with H19 are coronary artery disease, gastric cancer, neural tube defects, kidney cancer, infertility, etc., which correspond to cardiovascular, cancer, neurological, renal, and reproduction, respectively.

- For some less studied lncRNAs on PubMed and LncRNADisease, the same process is executed. The potential disease association results of APF are shown in Fig. 3. Here, APF lncRNA was selected with only 2 reports in PubMed and only associated with myocardial infarction disease in the LncRNADisease database. By conducting the above process, more disease associations are inferred, including diabetes mellitus type 2, metabolic syndrome X, obesity, asthma and hair diseases. These unreported diseases can provide insightful directions for future work with biological and medical researchers.

All the sequences of lncRNAs, proteins, predicted interacting proteins and inferred potential disease associations for each lncRNA can be downloaded at our website (<http://csbg-jlu.site/lpc/download>). This case study demonstrates the effectiveness and practicability of our Capsule-LPI tool. Furthermore, with the interacting proteins





predicted by Capsule-LPI, we can not only analyse their association with disease but also gain more information inference for lncRNAs such as function, evolution and subcellular localization.

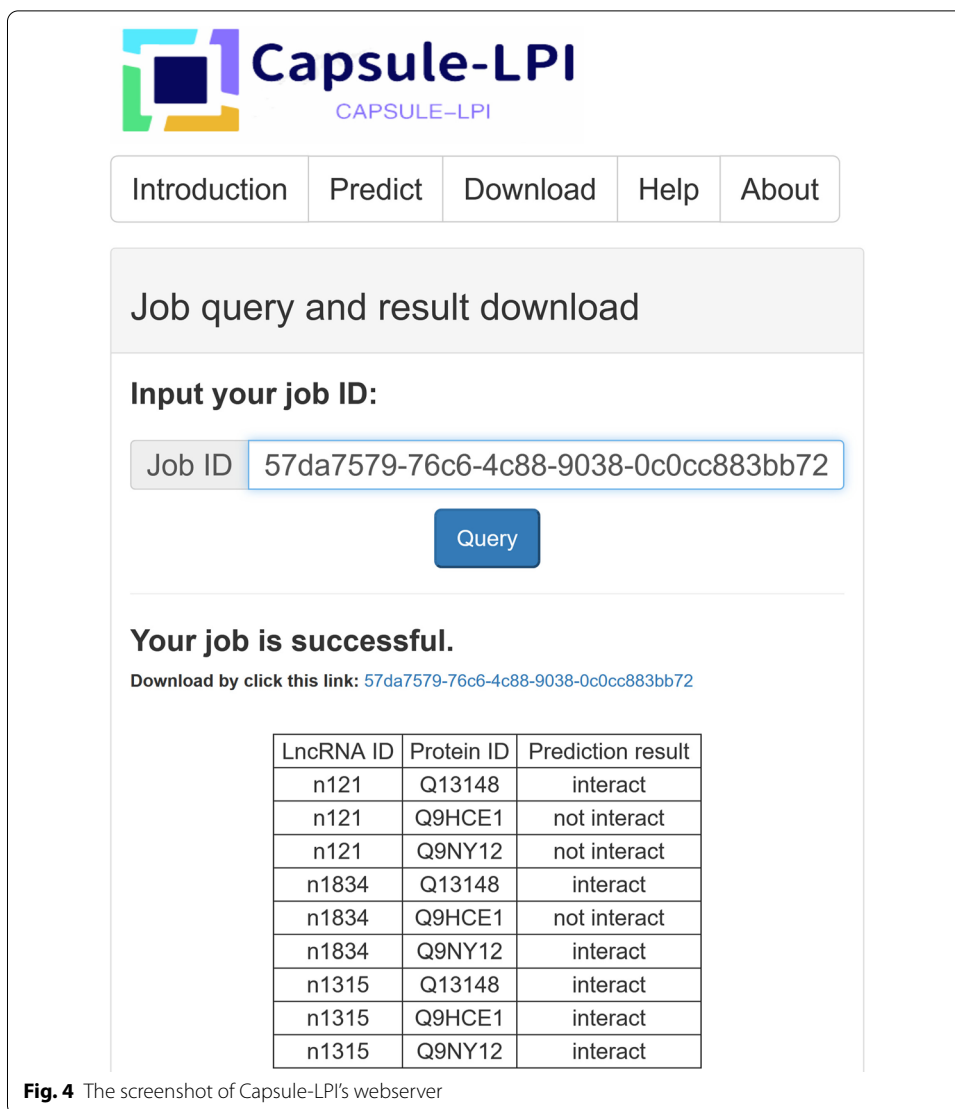
Webserver of capsule-LPI

We developed a webserver of Capsule-LPI with a user-friendly interface to facilitate users. The screenshot of the webserver of Capsule-LPI is shown in Fig. 4. Capsule-LPI allows online prediction of large volumes of data. The upper input limit for both lncRNA and protein sequences is 100, which means 10000 lncRNA–protein pairs can be predicted online. For each task submitted, a job ID can be assigned, and the prediction results can be downloaded by this job ID. For more functions and help, please refer to “help” on the website.

Conclusions

Identifying LPI is key to understanding lncRNA functions. Compared to experimental methods, computational methods are much more economical and efficient. Although some LPI prediction computational methods have been developed, how to integrate multimodal features from more perspectives and build deep learning architectures with better recognition performance has always been a challenging research highlight. Inspired by the better theory and the improved performance of the capsule network than CNN acting in image recognition, we propose a novel multichannel capsule network framework (Capsule-LPI) to integrate multimodal features for LPI prediction.

Capsule-LPI integrates four groups of multimodal features, including sequence features, motif information, physicochemical properties and secondary structure features. The architecture of Capsule-LPI is composed of four feature learning subnetworks and one capsule subnetwork. The multichannel framework of Capsule-LPI can make it better to integrate and learn multiple features by considering not only the



predictive tendencies of each feature but also other interaction information and the relationships between different features.

To comprehensively evaluate the performance of Capsule-LPI, three different kinds of experiments were conducted: (i) tool architecture comparison; (ii) evaluation of different feature combinations; and (iii) comparison with existing tools. Through comprehensive experimental comparisons and evaluations, we demonstrated that both multimodal features and the architecture of a multichannel capsule network can significantly improve the performance of LPI prediction. Capsule-LPI performs better than existing state-of-the-art tools. A webserver (<http://csbg-jlu.site/lpc/predict>) has been developed to be convenient for users.

This study provides a novel and feasible LPI prediction tool based on the integration of multimodal features and a capsule network. In the future, we expect to integrate more advanced structural features and external association information to further

improve the accuracy of Capsule-LPI and provide more convenient and practical tools for researchers.

Abbreviations

lncRNAs: Long noncoding RNAs; LPIs: LncRNA–protein interactions; LPI: LncRNA–protein interaction; SVM: Support vector machine; RF: Random forests; DNN: Deep neural network; FC: Fully connected network.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12859-021-04171-y>.

Additional file 1. Supplementary materials for Capsule-LPI Contains: S1. Dataset for Capsule-LPI; S2. The details of motifs; S3. The architecture of Capsule-LPI; S4. Features of LncADeep used in architecture comparison section; S5. Samples of the architectures of Capsule-LPI for different number of feature combination; S6. The potential lncRNA-disease association results.

Acknowledgements

The author thanks Jilin University for using high-performance computing facility and related services to support this work. The authors thank all the participators and providers of online lncRNA and protein resources. The authors thank all contributors to open source software. The authors also thank the editor and anonymous reviewers for handling this manuscript.

Authors' contributions

YL conceived the study and participated in its design and coordination, and drafted the manuscript. HS participated in the design, carried out the model analysis, design the webserver and drafted the manuscript. SYF helped to draft the manuscript. WD and QZ helped to build the webserver. SYH helped to extract the features of lncRNA and proteins. All authors read and approved the final manuscript.

Funding

This work was supported by the National Natural Science Foundation of China (61872418), Natural Science Foundation of Jilin Province (20180101331JC and 20180101050JC). The funding bodies have not played any role in the design of the study, the collection, analysis, interpretation of data, or the writing of the manuscript.

Availability of data and materials

The source code of Capsule-LPI and datasets can be accessed at the following URL: <http://csbg-jlu.site/lpc/download/>.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Key Laboratory of Symbolic Computation and Knowledge Engineering, Ministry of Education, College of Computer Science and Technology, Jilin University, Qianjin Street, 130012 Changchun, China. ²Department of Computer Science, Faculty of Engineering, University of Bristol, Bristol BS8 1UB, UK.

Received: 22 September 2020 Accepted: 5 May 2021

Published online: 13 May 2021

References

1. Gutschner T, Diederichs S. The hallmarks of cancer: A long non-coding rna point of view. *RNA Biology*. 2012;9:703–19.
2. Guttman M, Rinn JL. Modular regulatory principles of large non-coding rnas. *Nature*. 2012;482:339–46.
3. Pang KC, Frith MC, Mattick JS. Rapid evolution of noncoding rnas: lack of conservation does not mean lack of function. *Trends Genet*. 2006;22:1–5.
4. Kutter C, Watt S, Stefflova K, Wilson MD, Goncalves A, Ponting CP, Odom DT, Marques AC. Rapid turnover of long noncoding rnas and the evolution of gene expression. *PLoS Genet*. 2012;8:1002841.
5. Kung JT, Colognori D, Lee JT. Long noncoding rnas: past, present, and future. *Genetics*. 2013;193:651–69.
6. Wilusz JE, Sunwoo H, Spector DL. Long noncoding rnas: functional surprises from the rna world. *Genes Dev*. 2009;23:1494–504.

7. Harries LW. Long non-coding rnas and human disease. *Biochem Soc Trans.* 2012;40:902–6.
8. Fu M, Zou C, Pan L, Liang W, Qian H, Xu W, Jiang P, Zhang X. Long noncoding rnas in digestive system cancers: Functional roles, molecular mechanisms, and clinical implications (review). *Oncol Rep.* 2016;36:1207–18.
9. Rathinasamy B, Velmurugan BK. Role of lincrnas in the cancer development and progression and their regulation by various phytochemicals. *Biomedicine & Pharmacotherapy.* 2018;102:242–8.
10. Dangelmaier, E., Lal, A.: Adaptor proteins in long noncoding rna biology. *Biochimica et Biophysica Acta (BBA)–Gene Regulatory Mechanisms* 1863, 194370 (2020)
11. McHugh, C., Russell, P., Guttman, M.: McHugh, ca, russell, p and guttman, m. methods for comprehensive experimental identification of rna-protein interactions. *genome biol* 15: 203. *Genome biology* 15, 203 (2014)
12. Muppirlala UK, Honavar VG, Dobbs DJBB. Predicting rna-protein interactions using only sequence information. 2011;12:1–11.
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 Genomics.* 2013;14:651.
14. Suresh V, Liu L, Adjero D, Zhou X. Rpi-pred: predicting ncRNA-protein interaction using sequence and structural information. *Nucleic Acids Research.* 2015;43:1370–9.
15. Akbaripour-Elahabad, M., Zahiri, J., Rafeh, R., Eslami, M., Azari, M.J.J.o.T.B.: rpicool: A tool for in silico rna-protein interaction detection using random forest **402**, 1–8 (2016)
16. 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;2015:671950.
17. Ge M, Li A, Wang M. A bipartite network-based method for prediction of long non-coding rna-protein interactions. *Genomics Proteomics Bioinformatics.* 2016;14:62–71.
18. 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.
19. 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;14:1006616.
20. 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.
21. 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.
22. Hu H, Zhang L, Ai H, Zhang H, Fan Y, Zhao Q, Liu H. Hlpi-ensemble: Prediction of human lncRNA-protein interactions based on ensemble strategy. *RNA Biol.* 2018;15:797–806.
23. Yi HC, You ZH, Cheng L, Zhou X, Jiang TH, Li X, Wang YB. Learning distributed representations of rna and protein sequences and its application for predicting lncRNA-protein interactions. *Comput Struct Biotechnol J.* 2020;18:20–6.
24. Pan X, Fan YX, Yan J, Shen HB. lpmminer: hidden ncRNA-protein interaction sequential pattern mining with stacked autoencoder for accurate computational prediction. *BMC Genomics.* 2016;17:582.
25. Cheng, Y., Yang, L., Man, Z., Xie, H., Zhang, C., Wang, M.D., Zhu, H.J.B.: Lncadeep: An ab initio lincrna identification and functional annotation tool based on deep learning, 22 (2018)
26. Zhang SW, Zhang XX, Fan XN, Li WN. Lpi-cnncp: Prediction of lncRNA-protein interactions by using convolutional neural network with the copy-padding trick. *Anal Biochem.* 2020;601:113767.
27. Zhang Y, Jia C, Kwok CK. Predicting the interaction biomolecule types for lincrna: an ensemble deep learning approach. *Brief Bioinform.* 2020.
28. Wekesa JS, Meng J, Luan Y. A deep learning model for plant lncRNA-protein interaction prediction with graph attention. *Mol Genet Genomics.* 2020;295:1091–102.
29. Wekesa JS, Meng J, Luan Y. Multi-feature fusion for deep learning to predict plant lncRNA-protein interaction. *Genomics.* 2020;112:2928–36.
30. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: *European Conference on Computer Vision*
31. Hinton GEJS. Deep belief networks. 2009;4:5947.
32. Williams R, Zipser DJNC. A learning algorithm for continually running fully recurrent neural networks. 2014;1:270–80.
33. Schuster, M., Paliwal, K.K.J.I.T.o.S.P.: Bidirectional recurrent neural networks **45**, 2673–2681 (2002)
34. Laar, P.v.d., Heskens, T., Gielen, S.J.N.N.: Task-dependent learning of attention **10**, 981–992 (1997)
35. Sabour, S., Frosst, N., Hinton, G.E.: Dynamic routing between capsules (2017)
36. Zhou, J., Cui, G., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., Sun, M.: Graph neural networks: A review of methods and applications (2018)
37. Deng Y, Xu X, Qiu Y, Xia J, Zhang W, Liu S. A multimodal deep learning framework for predicting drug-drug interaction events. *Bioinformatics.* 2020;36:4316–22.
38. Yuan, J., Wu, W., Xie, C., Zhao, G., Zhao, Y., Chen, R.: Npinter v2.0: an updated database of ncRNA interactions. *Nucleic Acids Research* 42, 104–108 (2013)
39. He H, Garcia EA. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering.* 2009;21:1263–84.
40. Liu XY, Wu J, Zhou ZH. Exploratory undersampling for class-imbalance learning. 2009;39.
41. 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. *Molecular Therapy–Nucleic Acids.* 2018;11:337–44.
42. Pan JF, Wang T, Yu YH, Zhang DB. Preparation and thermal properties of non-equilibrium al/ptfe reactive materials. *Hanneng Cailiao/Chinese Journal of Energetic Materials.* 2016;24:582–6.
43. Shen J, Zhang J, Luo X, Zhu W, Yu K, Chen K, Li Y, Jiang H. Predicting protein-protein interactions based only on sequences information. *Proc Natl Acad Sci U S A.* 2007;104:4337–41.
44. Jiang, P., Singh, M., Collier, H.A., Zavolan, M.J.P.C.B.: Computational assessment of the cooperativity between rna binding proteins and micrnas in transcript decay **9**, 1003075 (2013)

45. Pancaldi V, Bähler J. In silico characterization and prediction of global protein-mrna interactions in yeast. *NUCLEIC ACIDS RES.* 2011;39.
46. Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, Gueroussov S, Albu M, Zheng H, Yang A, Na H, Irimia M, Matzat LH, Dale RK, Smith SA, Yarosh CA, Kelly SM, Nabet B, Mecnas D, Li W, Laishram RS, Qiao M, Lipshitz HD, Piano F, Corbett AH, Carstens RP, Frey BJ, Anderson RA, Lynch KW, Penalva LOF, Lei EP, Fraser AG, Blencowe BJ, Morris QD, Hughes TR. A compendium of rna-binding motifs for decoding gene regulation. *Nature.* 2013;499:172–7.
47. Morozova N, Allers J, Myers J, Shamoo YJB. Protein-rna interactions: exploring binding patterns with a three-dimensional superposition analysis of high resolution structures. 2006;22:2746–52.
48. Bull HB, Breese K. Surface tension of amino acid solutions: A hydrophobicity scale of the amino acid residues. *Archives of Biochemistry and Biophysics.* 1974;161:665–70.
49. Kyte, J., Doolittle, R.F.J.J.o.M.B.: A simple method for displaying the hydropathic character of a protein **157**, 105–132 (1982)
50. Zimmerman, J.M., Eliezer, N., Simha, R.J.J.o.T.B.: The characterization of amino acid sequences in proteins by statistical methods **21**, 170–201 (1968)
51. Grantham R. Amino acid difference formula to help explain protein evolution. *Science.* 1974;185:862–4.
52. Eisenberg, D., Schwarz, E., Komaromy, M., Wall, R.J.J.o.M.B.: Analysis of membrane and surface protein sequences with the hydrophobic moment plot **179**, 125–142 (1984)
53. Hopp TP, Woods KR. Prediction of protein antigenic determinants from amino acid sequences. 1981;78:3824–8.
54. Lorenz, R., Bernhart, S.H., Höner zu Siederdisen, C., Tafer, H., Flamm, C., Stadler, P.F., Hofacker, I.L.: ViennaRNA package 2.0. *Algorithms for Molecular Biology* 6, 26 (2011)
55. Frishman D, Argos P. Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence. *Protein Engineering, Design and Selection.* 1996;9:133–42.
56. Chou, P.Y., Fasman, G.D.J.A.i.E., Biology, R.A.o.M.: Prediction of the secondary structure of proteins from their amino acid sequence **47**, 145–148 (1978)
57. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research.* 2014;15:1929–58.

Publisher's Note

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

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

