LncRNA-protein interactions play important roles in post-transcriptional gene regulation, polyadenylation, splicing and translation. Identification of lncRNA-protein interactions helps to understand lncRNA-related activities. Existing computational methods utilize multiple lncRNA features or multiple protein features to predict lncRNA-protein interactions, but features are not available for all lncRNAs or proteins; most of existing methods are not capable of predicting interacting proteins (or lncRNAs) for new lncRNAs (or proteins), which don’t have known interactions. In this paper, we propose the sequence-based feature projection ensemble learning method, “SFPEL-LPI”, to predict lncRNA-protein interactions. First, SFPEL-LPI extracts lncRNA sequence-based features and protein sequence-based features. Second, SFPEL-LPI calculates multiple lncRNA-lncRNA similarities and protein-protein similarities by using lncRNA sequences, protein sequences and known lncRNA-protein interactions. Then, SFPEL-LPI combines multiple similarities and multiple features with a feature projection ensemble learning frame. In computational experiments, SFPEL-LPI accurately predicts lncRNA-protein associations and outperforms other state-of-the-art methods. More importantly, SFPEL-LPI can be applied to new lncRNAs (or proteins). The case studies demonstrate that our method can find out novel lncRNA-protein interactions, which are confirmed by literature. Finally, we construct a user-friendly web server, available at http://www.bioinfotech.cn/SFPEL-LPI/.

Abstract

LncRNA-protein interactions play important roles in post-transcriptional gene regulation, polyadenylation, splicing and translation. Identification of lncRNA-protein interactions helps to understand lncRNA-related activities. Existing computational methods utilize multiple lncRNA features or multiple protein features to predict lncRNA-protein interactions, but features are not available for all lncRNAs or proteins; most of existing methods are not capable of predicting interacting proteins (or lncRNAs) for new lncRNAs (or proteins), which don’t have known interactions. In this paper, we propose the sequence-based feature projection ensemble learning method, “SFPEL-LPI”, to predict lncRNA-protein interactions. First, SFPEL-LPI extracts lncRNA sequence-based features and protein sequence-based features. Second, SFPEL-LPI calculates multiple lncRNA-lncRNA similarities and protein-protein similarities by using lncRNA sequences, protein sequences and known lncRNA-protein interactions. Then, SFPEL-LPI combines multiple similarities and multiple features with a feature projection ensemble learning frame. In computational experiments, SFPEL-LPI accurately predicts lncRNA-protein associations and outperforms other state-of-the-art methods. More importantly, SFPEL-LPI can be applied to new lncRNAs (or proteins). The case studies demonstrate that our method can find out novel lncRNA-protein interactions, which are confirmed by literature. Finally, we construct a user-friendly web server, available at http://www.bioinfotech.cn/SFPEL-LPI/.

Author summary

LncRNA-protein interactions play important roles in post-transcriptional gene regulation, polyadenylation, splicing and translation. Identification of lncRNA-protein interactions helps to
In recent years, many computational methods have been proposed to predict IncRNA-protein interactions, in order to screen IncRNA-protein interactions and guide wet experiments. There are two types of computational methods: binary classification methods and semi-supervised learning methods. The binary classification methods take known interacting IncRNA-protein pairs as positive instances and non-interacting pairs as negative instances, and build binary classification-based models. Muppirala et al. [11] used RNA-protein interactions as positive instances, and randomly selected twice number of protein-RNA pairs without interaction information as negative samples, and then built prediction models by using naive Bayes. Suresh et al. [12] proposed a support vector machine-based predictor “RPI-Pred” to predict protein-RNA interactions based on their sequences and structures. Xiao et al. [13] used the HeteSim measure to score IncRNA-protein pairs, and then built an SVM classifier based on HeteSim scores. However, binary classification-based methods are influenced by the imbalance ratio between positive instances and negative instances, and how to select high-quality negative instances is challenging. Semi-supervised learning methods formulate the IncRNA-protein interaction prediction as semi-supervised learning tasks. Lu et al. [14] used matrix multiplication to score each RNA-protein pair for prediction. Li et al. [15] proposed a heterogeneous network-based method “LPI-HN”, which integrated the IncRNA-IncRNA similarity network, the IncRNA-protein interaction network and the protein-protein interaction network. Then, a random walk with restart was implemented on the heterogeneous network to infer IncRNA-protein interactions. Yang et al. [16] proposed the HeteSim algorithm, which can predict IncRNA-protein relation based on the heterogeneous IncRNA-protein network. Ge et al. [17] proposed a computational method “LPBIN” based on the

Existing computational methods utilize diverse IncRNA features and protein features, but features are not available for all IncRNAs or proteins, and these methods cannot work when information is unavailable. In addition, many IncRNAs (or proteins) don’t have known interactions with any protein (or IncRNA), and we name them as new IncRNAs (or proteins). Most existing methods are not capable of predicting interacting proteins (or IncRNAs) for new IncRNAs (or proteins).

In this paper, we propose the sequence-based feature projection ensemble learning method, “SFPEL-LPI”, to predict IncRNA-protein interactions. First, SFPEL-LPI extracts IncRNA sequence-based features and protein sequence-based features. Second, SFPEL-LPI calculates multiple IncRNA-IncRNA similarities and protein-protein similarities by using IncRNA sequences, protein sequences and known IncRNA-protein interactions. Then, SFPEL-LPI combines multiple similarities and multiple features with a feature projection ensemble learning frame. Computational experiments demonstrate that SFPEL-LPI predicts IncRNA-protein associations accurately and outperforms other state-of-the-art methods. More importantly, SFPEL-LPI can be applied to new IncRNAs (or proteins). The case studies demonstrate that our method can find out novel IncRNA-protein interactions.

Materials and methods

Dataset

Several databases facilitate the IncRNA-protein interaction prediction. NPInter database [24] includes experimental interactions among non-coding RNA and biomolecules (i.e. proteins, genomic DNAs and RNAs). NONCODE is an integrated information resource for non-coding RNAs. SUPERFAMILY [25] is a database of structural and functional annotation for all proteins and genomes. As far as we know, IncRNA-protein interactions from NPInter v2.0 database were widely used in related studies [20, 21, 26–29]. Based on NPInter v2.0 interactions, we compiled a dataset containing 4158 IncRNA-protein interactions between 990 IncRNAs and 27 proteins. Moreover, we collected the sequences of these IncRNAs and proteins from NONCODE and SUPERFAMILY respectively. We adopt NPInter v2.0 dataset as the benchmark dataset to test the performances of prediction models.

Here, we introduce notations about the dataset. Given a set of IncRNAs \( \mathcal{I} \) and a set of proteins \( \mathcal{P} \), known IncRNA-protein interactions can be represented by an \( s \times t \) interaction matrix \( Y \), where \( Y_{ij} = 1 \) if the IncRNA \( L_i \) interacts with the protein \( P_j \), otherwise \( Y_{ij} = 0 \).

Features for IncRNAs and proteins

In this section, we describe two IncRNA features and two protein features, based on IncRNA sequences, protein sequences and known IncRNA-protein interactions. On one hand, a great number of features [30–36] can be extracted from IncRNAs sequences and proteins sequences, and feature-extraction tools such as Pse-in-One[37], BioSeq-Analysis[38], repRNA[39][40], imiRNA-PseDPC [41] and UltraPse [42] have been available. One the other hand, known IncRNA-protein interactions can bring features to describe IncRNAs and proteins.

LncRNA features.

The pseudo dinucleotide composition (PseDNC) [43–46] describes the contiguous local sequence-order information and the global sequence-order information of IncRNAs. The pseudo dinucleotide composition has several variants, and we use the parallel correlation pseudo dinucleotide composition, which contains the occurrences of different dinucleotides and the physicochemical properties of dinucleotides. The PseDNC feature vector of an RNA sequence \( L \) is defined as:

\[
\text{PseDNC}(L) = \left( f_1, f_2, \ldots, f_k \right)
\]

where \( f_k \) is the normalized occurrence frequency of dinucleotide in the RNA sequence \( L \); the parameter \( r \) is an integer, representing the highest counted rank of the correlation along an RNA sequence; \( w \) is the weight factor ranging from 0 to 1; \( \theta_j \) is the \( j \)-tier correlation factor reflecting the sequence-order correlation between all the \( j \)-th most contiguous dinucleotides along an RNA sequence. We obtain PseDNC feature vectors of IncRNAs by using the python package “repDNA”, and more details about PseDNC are described in [40].
Moreover, we define the interaction profiles (IP) of lncRNAs based on known lncRNA-protein interactions. For a lncRNA \( L \), its interaction profile is a binary vector encoding the presence or absence of interactions with every protein, denoted as \( \text{IP}_L \). Actually, the interaction profile of a lncRNA corresponds to a row vector of the interaction matrix \( Y \).

**Protein features.**

The pseudo amino acid composition (PseAAC) [47–49] describes the amino acid composition and the sequence-order information of proteins, and has been widely used for tasks in bioinformatics. PseAAC contains 20 components reflecting the occurrence frequency of amino acids in a protein as well as the additional factors reflecting sequence-order information. Thus, we use PseAAC as a feature to represent proteins. There are several variants of PseAAC, and we adopt the parallel correlation pseudo amino acid composition. The PseAAC feature vector of a protein sequence \( P \) is defined as:

where

\[ f_i \] is the normalized occurrence frequency of the 20 amino acids in the protein sequence \( P \); the parameter \( \tau \) is an integer, representing the highest counted rank of the correlation along a protein sequence; \( w \) is the weight factor ranging from 0 to 1; \( \theta \) is the \( j \)-tier correlation factor reflecting the sequence-order correlation between all the \( j \)-th most contiguous residues along a protein sequence. We obtain the PseAAC feature vectors of proteins by using web server “Pse-in-One”, and more details are described in [37].

Similar to the lncRNA interaction profiles, the protein interaction profile (IP) of a protein \( P \) is a binary vector specifying the presence or absence of interactions with every lncRNAs, denoted as \( \text{IP}_P \). The interaction profile of a protein corresponds to a column vector of the interaction matrix \( Y \).

**Similarities for lncRNAs and proteins**

In this section, we describe three lncRNA-lncRNA similarities and three protein-protein similarities.

**LncRNA-lncRNA similarities.**

As introduced in Section “LncRNA features”, we have two lncRNA features: PseDNC and IP, and thus use them to calculate two types of lncRNA-lncRNA similarities. There are different approaches to calculate similarity based on feature vectors, such as Jaccard similarity, Gauss similarity and cosine similarity. Here, we adopt the linear neighborhood similarity (LNS), which has been proposed in our previous work and successfully applied to many bioinformatics problems [21,34,50].

Moreover, we define the Smith Waterman subgraph similarity (SWSS) for lncRNAs. Smith Waterman algorithm [51] is a powerful tool to calculate similarity between biological sequences, but Smith Waterman algorithm only takes the sequence information into account. By considering sequence information and interactions information, we define Smith Waterman subgraph similarity (SWSS) between lncRNA \( L_i \) and lncRNA \( L_j \) as,

where

\[ \text{SW}(P_{\text{P}1}, P_{\text{P}2}) \] is the Smith Waterman score between protein \( P_{\text{P}1} \) and protein \( P_{\text{P}2} \); \( A(L_i) \) and \( A(L_j) \) are the set of proteins which interact with \( L_i \) and \( L_j \); \( n_1 = |A(L_i)| \) and \( n_2 = |A(L_j)| \).

Therefore, we obtain three lncRNA-lncRNA similarities: PseDNC similarity, IP similarity and SWSS similarity.

**Protein-protein similarities.**

As introduced in Section “Protein features”, we have two proteins features: PseAAC and IP. We also calculate two types of similarities by using the linear neighborhood similarity measure. Similarly, we can calculate the Smith Waterman Subgraph Similarity (SWSS) between two proteins \( P_i \) and \( P_j \)

where

\[ \text{SW}(P_{\text{P}1}, P_{\text{P}2}) \] is the Smith Waterman score between lncRNA \( L_{\text{P}1} \) and lncRNA \( L_{\text{P}2} \); \( A(P_i) \) and \( A(P_j) \) are the set of lncRNAs which interact with protein \( P_i \) and protein \( P_j \); \( m_1 = |A(P_i)| \) and \( m_2 = |A(P_j)| \).
Therefore, we obtain three protein-protein similarities: PseAAC similarity, IP similarity and SWSS similarity.

Feature projection ensemble learning method

Combining various features or fusing various features can usually lead to high-accuracy models [52–58]. We have \( n \) features for IncRNAs (or proteins), denoted as \( n \) feature matrices \( A_j \), and have \( m \) types of similarities for IncRNAs (or proteins), denoted as \( m \) similarity matrices \( S_{ij} \). The predicted IncRNA-protein interaction matrix is denoted as \( R \). The known IncRNA-protein interaction matrix is denoted as \( Y \). The flowchart of the feature projection ensemble learning method SFPEL-LPI is shown in Fig 1.

\[
\text{Download:} \quad \text{Fig 1. The flowchart of SFPEL-LPI for predicting IncRNA-protein interactions.}
\]

Objective function.

First, IncRNA (or protein) feature matrices \( A_j \) are respectively projected to the predicted IncRNA-protein interaction matrix \( R \) by using the projection matrices \( G_j \). We estimate the projection matrices \( G_j \) for features by minimizing the squared error between their products and the predicted IncRNA-protein interaction matrix \( R \). So we have:

\[
\| AG - R \|^2
\]

where \( \| \cdot \|_F \) is the Frobenius norm, and the projection matrices \( G_j \) are required to be nonnegative.

Then, we introduce the \( \ell_1 \)-norm regularization term of \( G_j \) to ensure the smoothness of the projection matrices. The predicted matrix \( R \) should be approximated to the known interaction matrix \( Y \). We can have

\[
\| R - Y \|^2 + \lambda \| G \|_1
\]

where \( \lambda \) is the regularization coefficient, and \( \mu \) is a trade-off parameter.

Local structure of data can be maintained effectively through constructing a weighted graph or a similarity graph on a scatter of data points. For example, Xu et al. [59] introduced the manifold regularization term to preserve the visual feature manifold structure. Nie et al. [60], Bai et al. [61], Cai et al. [62,63] adopted graph Laplacian matrix to keep the graph’s local structure. Moreover, the Studies [34,64–67] revealed that the combination of multiple similarities helps to improve performances. Inspired by pioneer work, we define a novel ensemble graph Laplacian regularization:

\[
\left\| D^{1/2} (W \odot \theta) D^{1/2} - Y \right\|^2 + \eta \sum_{i=1}^{m} \left\| D^{1/2} \theta_i D^{1/2} - \right\|^2
\]

where \( D \) is a diagonal matrix whose diagonal elements are corresponding row sums of \( W \), and \( \theta = [\theta_1, \theta_2, \ldots, \theta_i, \ldots, \theta_m] \) is a weight vector which is introduced to control the contribution of different graph Laplacian regularizations, and \( tr(\cdot) \) is the trace of a matrix. \( \eta > 1 \) is the exponent of \( \theta \), which ensures that all graph Laplacian regularizations contribute effectively for the maintaining of graph local structures.

By combining (4) and (5), we obtain the objective function of SFPEL-LPI:

\[
\text{minimize} \quad \| R - Y \|^2 + \lambda \| G \|_1 + \eta \sum_{i=1}^{m} \left\| D^{1/2} \theta_i D^{1/2} - \right\|^2
\]

We introduce the Lagrangian function (Lf) to solve the optimization problem in (6).

We calculate the partial derivatives of above function with respect to \( R, G_i \) and \( \theta_i \) and obtain the update rules about \( R, \theta_i \) and \( G_i \) (proof and deduction are provided in S1 File):
where $e$ is a column vector with all elements equal to 1, and has the same column dimensions as $X$. $\odot$ denotes element-wise multiplication (also well known as Hadamard product), and the division in (9) is element-wise division. We separate the positive and negative parts of matrix $A$ as

Thus, we update $R$, $G_l$ and $\theta_l$ based on (7), (8) and (9) alternatively until convergence.

Algorithms.

Following the method proposed in the Section “Objective function”, SFPEL-LPI can predict unobserved interactions between known lncRNAs and proteins. First, based on the lncRNA’s features, similarities and lncRNA-protein interactions, the prediction matrix $R_l$ could be obtained. Similarly, using protein’s features, similarities and protein-lncRNA interactions, the prediction matrix $R_p$ could be calculated. Then, SFPEL-LPI integrates the predictions based on lncRNAs and proteins as $M = (R_l + (R_p)^T)/2$. Therefore, the unobserved interactions are scored in the corresponding entries of $M$. Algorithm 1 describes how SFPEL-LPI predicts unobserved associations between known lncRNAs and known proteins.

In addition, SFPEL-LPI could also be applied to predict proteins (or lncRNAs) interacting with new lncRNAs (or proteins). After using Algorithm 1 to train the model, the projection matrix and the weighting parameters of lncRNA’s features as well as protein’s features: $G_{lu}$, $G_{lv}$, $\theta_{lu}$ and $\theta_{lv}$ could be obtained. Then, we can use the features of new lncRNAs (or proteins) and the trained parameters to predict their predictions. Algorithm 2 describes how SFPEL-LPI finishes this task.

Algorithm 1: Predicting unobserved associations between known lncRNAs and known proteins by SFPEL-LPI.

**Input:** observed lncRNA-protein interaction matrix, $Y_l$; observed protein-lncRNA interaction matrix, $Y_p = Y_l^T$; lncRNA feature matrices, $\{X_{l1}, X_{l2}, ..., X_{ln_l}\}$; protein feature matrices, $\{X_{p1}, X_{p2}, ..., X_{pm}\}$; lncRNA normalized similarity matrices, $\{W_{lu}, W_{l2}, ..., W_{ln_l}\}$; protein normalized similarity matrices, $\{W_{p1}, W_{p2}, ..., W_{pm}\}$; regularization parameter, $\mu > 0, \lambda > 0$; exponent parameter, $\eta > 1$.

**Output:** lncRNA-protein interaction prediction matrix, $M$; predicted lncRNA-protein interaction matrix, $R_l$; predicted protein-lncRNA interaction matrix, $R_p$; projection matrices of lncRNA features $\{G_{lu}, G_{l2}, ..., G_{ln_l}\}$; projection matrices of protein features $\{G_{p1}, G_{p2}, ..., G_{pm}\}$; weighting parameters of lncRNA similarity matrices, $\{\theta_{lu}, \theta_{l2}, ..., \theta_{ln_l}\}$; weighting parameters of protein similarity matrices, $\{\theta_{p1}, \theta_{p2}, ..., \theta_{pm}\}$.

**Initialize:**

for each $i (1 \leq i \leq n_l)$

initialize $G_{lu}$, $G_{p_l}$ with random values on interval $[0,1]$;

end for

for each $i (1 \leq i \leq m)$

initialize $\theta_{lu}$, $\theta_{p_l}$ as $1/m$;

end for

repeat

update $R_l$ via (7) with fixing $\alpha$;

for each $i (1 \leq i \leq n_l)$

update $G_{lu}$ via (8) with fixing $R_l$;

end for

for each $i (1 \leq i \leq m)$

update $\theta_{lu}$ via (9) with fixing $R_l$;

end for

Algorithm 2: Predicting unobserved interactions between known lncRNAs and new proteins by SFPEL-LPI.
the best AUPR score of 0.473 when implement five-fold cross validation.

To test influence of parameters, we consider all combinations of parameters

\[
\begin{align*}
\mu & \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^0, 10^1, 10^2, 10^3\}, \\
\eta & \in \{2^2, 2^3, 2^4, 2^5, 2^6, 2^7, 2^8\}.
\end{align*}
\]

We build SFPEL-LPI models by using different parameters, and implement five-fold cross validation to evaluate SFPEL-LPI models. SFPEL-LPI produces the best AUPR score of 0.473 when \( \mu = 10^{-3} \), \( \lambda = 10^{-4} \) \text{ and } \eta = 2^5. \) Then, we fix the parameter \( \eta = 2^5 \), and evaluate the influence of \( \mu \) and \( \lambda \). As shown in Fig 2A, \( \mu \) greatly influences the
performance of SFPEL-LPI, and a smaller value for $\mu$ is likely to produce better result. Further, we fix the parameters $\mu = 10^{-3}$ and $\lambda = 10^{-4}$ and test the influence of $\eta$. As illustrated in Fig 2B, the performances of SFPEL-LPI decrease as $\eta$ increases, and then remain unchanged after a threshold.

![Fig 2. The influence of parameters on AUPR of models.](https://doi.org/10.1371/journal.pcbi.1006616.g002)

The parameter $\eta$ is the index of similarity weights, and could control the relative contributions of different similarities. When fixing $\mu = 10^{-3}$ and $\lambda = 10^{-4}$, we analyze the relation between $\eta$ and lncRNA similarity measures $\theta_{\text{lncRNA}}$ (or protein similarity measures $\theta_{\text{protein}}$). As shown in Fig 3, similarities usually make different contributions to SFPEL-LPI models, and interaction profile similarities usually make more contributions than other similarities. With increase of $\eta$, different similarities are likely to make equal contributions.

Based on above discussion, we adopt $\mu = 10^{-3}, \lambda = 10^{-4}$ and $\eta = 2^2$ for SFPEL-LPI in the following studies.

![Fig 3.](https://doi.org/10.1371/journal.pcbi.1006616.g003)

### Performances of SFPEL-LPI

SFPEL-LPI can predict unobserved lncRNA-protein interactions between known lncRNAs and known proteins, and also can make predictions for new lncRNAs (or proteins). For different tasks, we adopt different evaluation schemes to split instances and implement five-fold cross validation under settings: $CV_{lp}$, $CV_{l}$ and $CV_{p}$.

Table 1 displays AUPR scores and AUC scores of SFPEL-LPI evaluated by $CV_{lp}$, $CV_{l}$ and $CV_{p}$. According to previous studies [70–72], a prediction model that can accurately recover the true interacting proteins (or lncRNAs) is usually desired and useful for the wet experimental validation. Thus, we calculate the proportion of correctly predicted true interactions at different top-ranked percentiles under $CV_{l}$ or $CV_{p}$. A new metric “recall @ top-ranked k %” is defined as the fraction of true interacting proteins (or lncRNAs) that are retrieved in the list of top-ranked k% predictions for a lncRNA (or protein). In Fig 4A, SFPEL-LPI performs effectively in predicting proteins (or lncRNAs) interacting with new lncRNAs (or proteins). The reason why the performances of predicting lncRNAs interacting with new proteins is not as well as the performances of predicting proteins interacting with new lncRNAs is that the number of lncRNAs (990) in our dataset is much more than the number of proteins (27). Consequently, less information is used to train SFPEL-LPI models.

![Fig 4.](https://doi.org/10.1371/journal.pcbi.1006616.g004)
interacting lncRNAs for selected new proteins.
https://doi.org/10.1371/journal.pcbi.1006616.g004

Table 1. Performances of SFPEL-LPI for predicting lncRNA-protein interactions.
https://doi.org/10.1371/journal.pcbi.1006616.t001

To further test capability of SFPEL-LPI for new proteins, we randomly select ten proteins to conduct experiments. In each experiment, a protein is used as the testing protein, and the model is constructed based on other proteins, all lncRNAs and their associations, and then predict lncRNAs interacting with the testing protein. AUC scores and AUPR scores are calculated based on the results for each protein. As shown in Fig 4B, SFPEL-LPI produces the AUPR values greater than 0.6 and the AUC values greater than 0.7 for most proteins, indicating great potential of predicting lncRNAs interacting with new proteins.

Comparison with state-of-the-art prediction methods

Several state-of-the-art computational methods have been proposed to predict lncRNA-protein interactions. Here, we adopt RWR[17], LPBNI[17], KATZLGO[19], LPI-ETSLP [20] and LPLNP [21] for comparison. RWR implemented random walk with restart to predict lncRNA-protein interactions. LPBNI constructed a lncRNA-protein bipartite network based on known lncRNA-protein interactions, and then predicted lncRNA-protein interactions by using the resource allocation algorithm. KATZLGO constructed a heterogeneous network based on lncRNA-lncRNA similarity, lncRNA-protein interactions and protein-protein similarity, and then adopted KATZ measure to calculate distances between lncRNAs and proteins in the network. LPI-ETSLP calculated lncRNA-lncRNA similarity and protein-protein similarity based on pairwise sequence Smith-Waterman scores, and then built semi-supervised link prediction classifier based on these similarities. LPLNP calculated three lncRNA-lncRNA similarities and two protein-protein similarities by using linear neighborhood similarity measure, and implemented label propagation to develop the integrated models.

First, we respectively build different prediction models based on the benchmark dataset. The benchmark methods were designed to predict unobserved interaction between know lncRNAs and know proteins. Therefore, we implement these methods and mainly evaluate their performances in predicting unobserved interactions under CV. As shown in Table 2, the AUPR values of RWR, LPBNI, KATZLGO, LPI-ETSLP, LPLNP and SFPEL-LPI are 0.236, 0.330, 0.286, 0.322, 0.459, 0.473, and AUC values are 0.850, 0.856, 0.760, 0.889, 0.910 and 0.920, respectively. SFPEL-LPI outperforms these five methods, and makes 100.4%, 43.3%, 65.4%, 46.9%, 3.1% improvements in terms of AUPR scores and 8.2%, 7.5%, 21.1%, 3.5%, 1.1% improvements in terms of AUC scores when compared with five benchmark methods. Though SFPEL-LPI produces slightly better performances than LPLNP in terms of AUPR and AUC, LPLNP utilizes more information than SFPEL-LPI for modeling. To be more specific, LPLNP uses three lncRNA features (“interaction profile”, “expression profile”, “sequence composition”) and two protein features (“interaction profile”, “CTD”), while SFPEL-LPI only used lncRNA sequences, protein lncRNAs and known lncRNA-protein interactions.

Table 2. Performances of prediction methods on the benchmark dataset.
https://doi.org/10.1371/journal.pcbi.1006616.t002

We conduct 20 runs of five-fold cross validation to evaluate methods, and take the paired t-test to analyze difference between SFPEL-LPI and benchmark methods. Table 3 demonstrates that SFPEL-LPI produces significantly better results than state-of-the-art methods in terms of AUC and AUPR.
The computational complexity is important for a computational method. To test the efficiency of SFPEL-LPI, we repeat 5-fold cross validation 20 times and compare running time of different methods on a PC with an Intel i7 7700k CPU and 16GB RAM. SFPEL-LPI costs a reasonable running time (29.42s) when compared with RWR (25.83s), LPBNI (4.01s), KATZLGO (4.36s), LPI-ETSLP (4.56s) and LPLNP (1337.64s).

Further, we randomly perturb all known IncRNA-protein interactions to test the robustness of prediction methods. To be more specific, we randomly remove 5% of known IncRNA-protein interactions and add the same number of nonexistent interactions, and then compile the perturbed dataset. We build different prediction models based on the perturbed dataset and evaluate their performances. Clearly, data perturbation brings noise, and decreases the performances of prediction models. As displayed in Fig 5, AUC scores of RWR, LPBNI, KATZLGO, LPI-ETSLP, LPLNP, SFPEL-LPI are 0.812, 0.820, 0.735, 0.865, 0.874 and 0.889; AUPR scores are 0.192, 0.268, 0.225, 0.271, 0.343 and 0.351. Although prediction models produce lower performances than that in Table 2, SFPEL-LPI still produces satisfying results, and outperforms RWR, LPBNI, KATZLGO, LPI-ETSLP and LPLNP.

Independent experiments

Here, we conduct independent experiments to evaluate the practical ability of SFPEL-LPI. As described in Section “Dataset”, NPInter v2.0 dataset was compiled from the V2.0 edition of NPInter database. NPInter database has been updated to V3.0 edition, and contains newly discovered IncRNA-protein interactions. Therefore, we train the prediction model based on the NPInter v2.0 dataset and predict new IncRNA-protein interactions, and then check up on predictions in the NPInter database. Fig 6 shows the number of confirmed interactions in top 20 predictions of all methods. Clearly, SFPEL-LPI finds out more interactions than benchmark methods. In addition, we observe that most of novel interactions identified by SFPEL-LPI have low ranks in the predictions of other benchmark methods, indicating that SFPEL-LPI can find out interactions ignored by these methods. Top predictions and their ranks are provided in S1 Table.

Web server

We develop a web server based on SFPEL-LPI to facilitate the IncRNA-protein interaction prediction, available at http://www.bioinfotech.cn/SFPEL-LPI/. Users can input IncRNA sequences (or protein sequences) or upload a text file with FASTA-formatted IncRNA sequences.
(or protein sequences) for prediction, and freely download the results and visualize the predicted lncRNA-protein interactions. Moreover, gene ontology (GO) terms of proteins are annotated for indicating lncRNAs' functions.

Fig 7 displays the top 10 predictions for the lncRNA "NONHSAT041930". "NONHSAT041930" named OIP5-AS1 (OIP5 antisense RNA 1), is a mammalian lncRNA that is abundant in the cytoplasm [73]. OIP5-AS1 has gained wide attention. In 2011, it was first identified to be involved in brain and eye development [74]. In 2016, Kim et al. [75] found that it can prevent HuR binding to target mRNAs and thus suppress the HuR-elicited proliferative phenotypes. Moreover, the lncRNA was found to interact with GAK mRNA, promoting GAK mRNA decay and hence reducing GAK protein levels and lowering cell proliferation [76]. Among top 10 predicted proteins interacting with OIP5-AS1, two proteins have already been known to have interactions with OIP5-AS1, which are included in the NPInter dataset. In addition, we find evidence from literature to support other six predicted proteins. For example, IG2B2P1, IG2B2P2, IG2B2P3, EWSR1 and TIAI have already been examined to interact with OIP5-AS1 according to lncRNA-protein interacting data report [77]. Protein Argonaute 2 (AGO2) is required for proper nuclear migration, pole cell formation, and cellularization during the early stages of embryonic development. Several studies [75, 78] showed that OIP5-AS1 is associated with AGO2. Moreover, annotated GO terms of predicted proteins indicate the function of the lncRNA OIP5-AS1: mRNA binding (GO: 0005845, GO: 0035925, GO: 0036002, GO: 0048027, GO: 0098808) and cell proliferation (GO:0022013). More details are provided in S2 Table. These encouraging instances demonstrate that the proposed method can successfully predict novel lncRNA-protein interactions.

Moreover, the server can predict interacting lncRNAs for proteins. For example, top 20 interacting lncRNAs of the protein "9606.ENSP00000240185" are shown in the Fig 8, and details are provided in S3 Table.

Discussion

This paper presents a novel lncRNA-protein interaction prediction method, namely sequence-based feature projection ensemble learning (SFPEL-LPI). The novelty of SFPEL-LPI comes from integrating sequence-derived features and similarities with a feature projection ensemble learning frame. Specifically, SFPEL-LPI only utilizes lncRNA sequences, protein sequences and known interactions to extract features, and calculates lncRNA-lncRNA similarities and protein-protein similarities. Since sequences are usually available for lncRNAs or proteins, SFPEL-LPI can make predictions for almost all lncRNA-protein pairs. Moreover, diverse information leads to the good performances of SFPEL-LPI.

To evaluate the performance of SFPEL-LPI, an extensive set of experiments were performed on the benchmark dataset under three CV setting: CV\(_p\), CV\(_l\) and CV\(_u\), compared with state-of-the-art lncRNA-protein interaction prediction methods. The promising results validate efficacy of the
proposed algorithm for predicting IncRNA-protein interactions, especially for the new IncRNAs or new proteins, which do not have known interactions. SFPEL-LPI outperforms five methods: RWR, LPBNI, KATZLGO, LPI-ETSLP, LPLNP, and makes 100.4%, 43.3%, 65.4%, 46.9%, 3.1% improvements in terms of AUPR scores. Further, we also analyze the running time of SFPEL-LPI and benchmark methods, and randomly perturb all known IncRNA-protein interactions to test the robustness of prediction methods. A web server is constructed to predict interacting proteins/IncRNAs for given IncRNAs/proteins. We adopt the IncRNA “NONHSAT041930” as an example to predict interacting proteins, and can find evidences to confirm novel IncRNA-protein interactions.

However, SFPEL-LPI still has several limitations. It has three parameters, and parameter tuning is time-consuming. In addition, known IncRNA-protein interactions are limited, and performances of SFPEL-LPI will be improved if more interactions are known.

Supporting information

S1 File. Proof and analysis of SFPEL-LPI.
https://doi.org/10.1371/journal.pcbi.1006616.s001
(PDF)

S2 File. The data of SFPEL-LPI.
https://doi.org/10.1371/journal.pcbi.1006616.s002
(MAT)

S1 Table. Top 20 predictions of SFPEL-LPI and their ranks in predictions of benchmark methods.
https://doi.org/10.1371/journal.pcbi.1006616.s003
(DOCX)

S2 Table. Top 10 interacting proteins of LncRNA “NONHSAT041930” (OIP5-AS1) predicted by SFPEL-LPI.
https://doi.org/10.1371/journal.pcbi.1006616.s004
(DOCX)

S3 Table. Top 20 interacting IncRNAs of protein “9606.ENSP00000240185” (TAR DNA-binding protein 43) predicted by SFPEL-LPI.
https://doi.org/10.1371/journal.pcbi.1006616.s005
(DOCX)

References

   View Article • PubMed/NCBI • Google Scholar

   View Article • PubMed/NCBI • Google Scholar

   View Article • PubMed/NCBI • Google Scholar

   View Article • PubMed/NCBI • Google Scholar

   View Article • PubMed/NCBI • Google Scholar

   View Article • PubMed/NCBI • Google Scholar

   View Article • PubMed/NCBI • Google Scholar

View Article • PubMed/NCBI • Google Scholar

View Article • PubMed/NCBI • Google Scholar

View Article • PubMed/NCBI • Google Scholar

View Article • Google Scholar

View Article • Google Scholar

View Article • Google Scholar

View Article • Google Scholar

View Article • PubMed/NCBI • Google Scholar

View Article • PubMed/NCBI • Google Scholar

View Article • PubMed/NCBI • Google Scholar

View Article • Google Scholar

View Article • PubMed/NCBI • Google Scholar

View Article • PubMed/NCBI • Google Scholar

View Article • Google Scholar
   View Article • PubMed/NCBI • Google Scholar

   View Article • Google Scholar

   View Article • PubMed/NCBI • Google Scholar

   View Article • PubMed/NCBI • Google Scholar

   View Article • PubMed/NCBI • Google Scholar

   View Article • Google Scholar

   View Article • PubMed/NCBI • Google Scholar

   View Article • PubMed/NCBI • Google Scholar

   View Article • PubMed/NCBI • Google Scholar

   View Article • Google Scholar

   View Article • PubMed/NCBI • Google Scholar

   View Article • PubMed/NCBI • Google Scholar

   View Article • PubMed/NCBI • Google Scholar

   View Article • PubMed/NCBI • Google Scholar

   View Article • PubMed/NCBI • Google Scholar


