PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

SFPEL-LPI: Sequence-based feature projection ensemble learning for predicting LncRNA-protein interactions

Wen Zhang[1,2]*, Xiang Yue[3], Guifeng Tang[2], Wenjian Wu[4], Feng Huang[2], Xining Zhang[2]*

1 College of Informatics, Huazhong Agricultural University, Wuhan, China, **2** School of Computer Science, Wuhan University, Wuhan, China, **3** Department of Computer Science and Engineering, The Ohio State University, Columbus, United States of America, **4** Electronic Information School, Wuhan University, Wuhan, China

* zhangwen@whu.edu.cn, zhangwen@mail.hzau.edu.cn (WZ); zhangxn@whu.edu.cn (XZ)

Citation: Zhang W, Yue X, Tang G, Wu W, Huang F, Zhang X (2018) SFPEL-LPI: Sequence-based feature projection ensemble learning for predicting LncRNA-protein interactions. PLoS Comput Biol 14 (12): e1006616. https://doi.org/10.1371/journal.pcbi.1006616

Editor: Ilya Ioshikhes, Ottawa University, CANADA

Received: June 19, 2018

Accepted: November 2, 2018

Published: December 11, 2018

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: WZ was supported by the National Natural Science Foundation of China (61772381, 61572368), the Fundamental Research Funds for the Central Universities (2042017kf0219, 2042018kf0249), Huazhong Agricultural University Scientific & Technological Self-innovation Foundation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

LncRNA-protein interactions play important roles in post-transcriptional gene regulation, poly-adenylation, 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, poly-adenylation, splicing and translation. Identification of lncRNA-protein interactions helps to understand lncRNA-related activities. In this paper, we propose a novel computational method "SFPEL-LPI" to predict lncRNA-protein interactions. SFPEL-LPI makes use of lncRNA sequences, protein sequences and known lncRNA-protein associations to extract features and calculate similarities for lncRNAs and proteins, and then

Competing interests: The authors have declared that no competing interests exist.

combines them with a feature projection ensemble learning frame. SFPEL-LPI can predict unobserved interactions between lncRNAs and proteins, and also can make predictions for new lncRNAs (or proteins), which have no interactions with any proteins (or lncRNAs). SFPEL-LPI produces high-accuracy performances on the benchmark dataset when evaluated by five-fold cross validation, and outperforms state-of-the-art methods. The case studies demonstrate that SFPEL-LPI can find out novel associations, which are confirmed by literature. To facilitate the lncRNA-protein interaction prediction, we develop a user-friendly web server, available at http://www.bioinfotech.cn/SFPEL-LPI/.

This is a *PLOS Computational Biology* Methods paper.

Introduction

Long noncoding RNAs (lncRNAs) are a class of transcribed RNA molecules with a length of more than 200 nucleotides that do not encode proteins [1,2]. Since lncRNAs are involved in important biological regulations [3–5], lncRNAs have gained widespread attention. Studies [5–9] revealed that lncRNAs can interact with proteins, and then activate post-transcriptional gene regulation, poly-adenylation, splicing and translation. Identification of lncRNA-protein interactions helps to understand lncRNAs' functions. There exist a large number of unexplored lncRNAs and proteins, which makes it impossible to examine their interactions efficiently and effectively through wet experiments.

In recent years, many computational methods have been proposed to predict lncRNA-protein interactions, in order to screen lncRNA-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 lncRNA-protein pairs as positive instances and non-interacting pairs as negative instances, and build binary classification-based models. Muppirala et al. [10] adopted the k-mer composition to encode RNA sequences and protein sequences, and used SVM and random forest to build prediction models. Wang 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 lncRNA-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 lncRNA-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 "LPIHN", which integrated the lncRNA-lncRNA similarity network, the lncRNA-protein interaction network and the protein-protein interaction network. Then, a random walk with restart was implemented on the heterogeneous network to infer lncRNA-protein interactions. Yang et al. [16] proposed the Hetesim algorithm, which can predict lncRNA-protein relation based on the heterogeneous lncRNA-protein network. Ge et al. [17] proposed a computational method "LPBNI" based on the lncRNA-protein bipartite network inference.

Zheng et al. [18] constructed multiple protein-protein similarity networks to predict lncRNA-protein interactions. Zhang et al. [19] employed KATZ measure to calculate similarities between lncRNAs and proteins in a global network, which were constructed based on lncRNA-lncRNA similarity, lncRNA-protein associations and protein-protein interactions. Hu et al. [20] presented the eigenvalue transformation-based semi-supervised link prediction method "LPI-ETSLP". Zhang et al. [21] proposed a linear neighborhood propagation method (LPLNP) by combining interaction profiles, expression profiles, sequence composition of lncRNAs and interaction profile, CTD feature of proteins. Moreover, there are related works about the DNA-protein binding prediction [22,23].

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

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. Computational experiments demonstrate that SFPEL-LPI predicts lncRNA-protein associations accurately 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.

Materials and methods

Dataset

Several databases facilitate the lncRNA-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, lncRNA-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 lncRNA-protein interactions between 990 lncRNAs and 27 proteins. Moreover, we collected the sequences of these lncRNAs 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 lncRNAs $\mathcal{L} = \{L_1, L_2, \cdots, L_s\}$ and a set of proteins $\mathcal{P} = \{P_1, P_2, \cdots, P_t\}$, known lncRNA-protein interactions can be represented by an $s \times t$ interaction matrix Y, where $Y_{ij} = 1$ if the lncRNA L_i interacts with the protein P_j, otherwise $Y_{ij} = 1$.

Features for lncRNAs and proteins

In this section, we describe two lncRNA features and two protein features, based on lncRNA sequences, protein sequences and known lncRNA-protein interactions. On one hand, a great number of features [30–36] can be extracted from lncRNAs 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 lncRNA-protein interactions can bring features to describe lncRNAs 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 lncRNAs. 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:

$$L = [d_1, d_2, \cdots, d_{16}, d_{16+1}, \cdots, d_{16+\tau}]$$

where

$$d_k = \begin{cases} \dfrac{f_k}{\sum_{i=1}^{16} f_i + w\sum_{j=1}^{\tau}\theta_j} 1 \le k \le 16 \\ \dfrac{w\theta_{k-16}}{\sum_{i=1}^{16} f_i + w\sum_{j=1}^{\tau}\theta_j} 17 \le k \le 16+\tau \end{cases}$$

where f_k is the normalized occurrence frequency of dinucleotide in the RNA sequence L; the parameter τ 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; θ_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 lncRNAs 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_i, its interaction profile is a binary vector encoding the presence or absence of interactions with every protein, denoted as IP_{L_i}. Actually, the interaction profile of a lncRNA corresponds to a row vector of the interaction matrix Y, $IP_{L_i} = Y(i,:)$.

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:

$$P = [x_1, x_2, \cdots, x_{20}, x_{20+1}, \cdots, x_{20+\tau}]$$

where

$$x_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{20} f_i + w\sum_{j=1}^{\tau}\theta_j} 1 \le \mathrm{u} \le 20 \\ \dfrac{w\theta_{u-20}}{\sum_{i=1}^{20} f_i + w\sum_{j=1}^{\tau}\theta_j} 21 \le \mathrm{u} \le 20+\tau \end{cases}$$

where f_i is the normalized occurrence frequency of the 20 amino acids in the protein sequence P; the parameter τ 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; θ_j 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_i is a binary vector specifying the presence or absence of interactions with every lncRNAs, denoted as IP_{p_i}. The interaction profile of a protein corresponds to a column vector of the interaction matrix Y, $IP_{p_i} = Y(:, i)$.

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,

$$SWSS\left(L_i, L_j\right) = \sum_{P_{o1} \in \mathrm{A}(L_i)} \sum_{P_{o2} \in \mathrm{A}(L_j)} \frac{SW(P_{o1}, P_{o2})}{n1 \times n2} \tag{1}$$

where $SW(P_{o1},P_{o2})$ is the Smith Waterman score between protein P_{o1} and protein P_{o2}. $\mathrm{A}(L_i)$ and $\mathrm{A}(L_j)$ are the set of proteins which interact with L_i and L_j. $n1 = |\mathrm{A}(L_i)|$ and $n2 = |\mathrm{A}(L_j)|$.

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,

$$SWSS\left(P_i, P_j\right) = \sum_{L_{o1} \in \mathrm{A}(P_i)} \sum_{L_{o2} \in \mathrm{A}(P_j)} \frac{SW(L_{o1}, L_{o2})}{m1 \times m2} \tag{2}$$

where $SW(P_{o1},P_{o2})$ is the Smith Waterman score between lncRNA L_{o1} and lncRNA L_{o2}. $\mathrm{A}(P_i)$ and $\mathrm{A}(P_j)$ are the set of lncRNAs which interact with protein P_i and protein P_j. $m1 = |\mathrm{A}(P_i)|$ and $m2 = |\mathrm{A}(P_j)|$.

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 lncRNAs (or proteins), denoted as n feature matrices $\{X_i\}_{i=1}^{n}$, and have m types of similarities for lncRNAs (or proteins), denoted as m similarity matrices $\{W_i\}_{i=1}^{m}$. The predicted lncRNA-protein interaction matrix is denoted as R. The

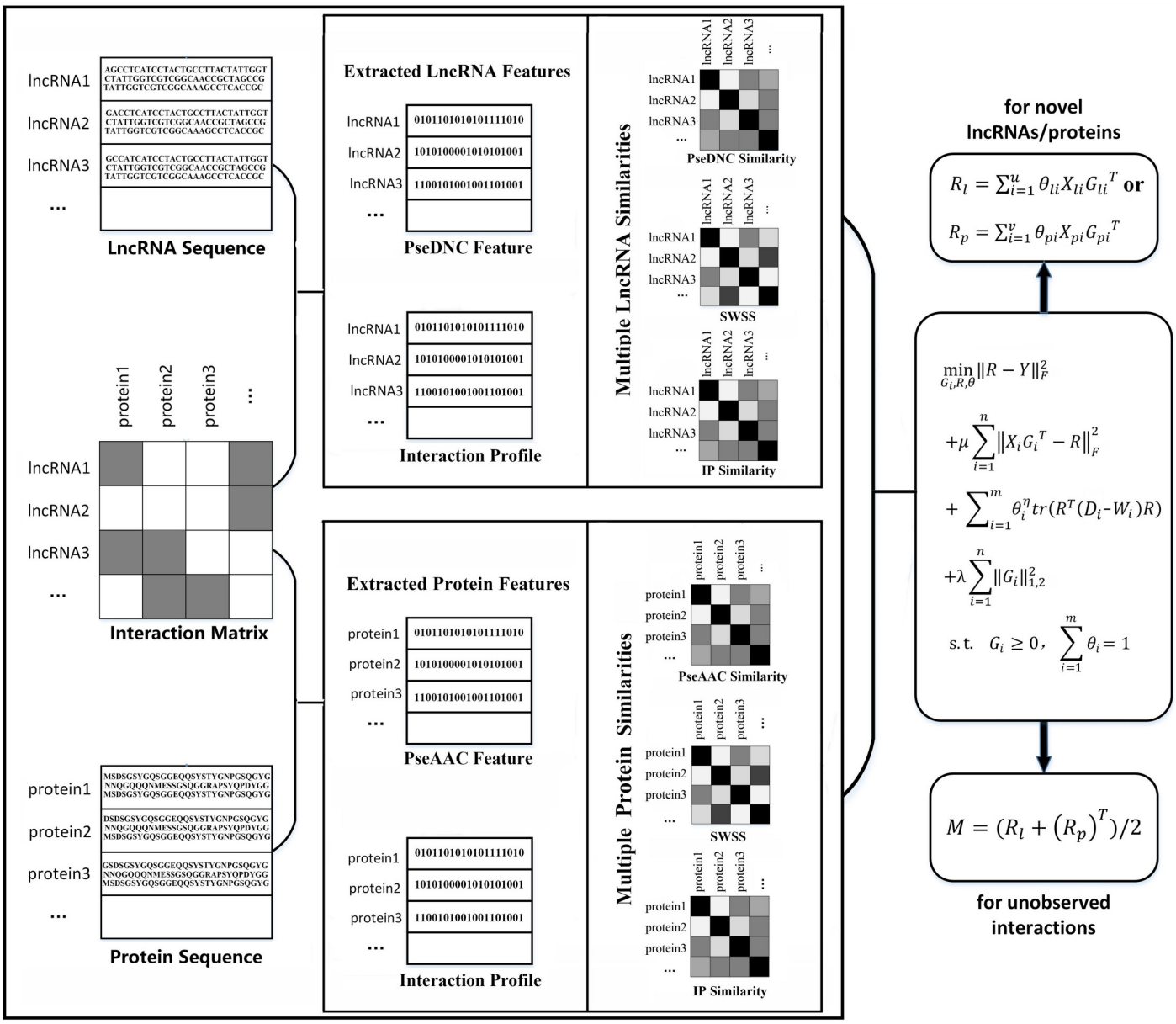

Fig 1. The flowchart of SFPEL-LPI for predicting lncRNA-protein interactions.

https://doi.org/10.1371/journal.pcbi.1006616.g001

known lncRNA-protein interaction matrix is denoted as Y. The flowchart of the feature projection ensemble learning method SFPEL-LPI is shown in Fig 1.

Objective function. First, lncRNA (or protein) feature matrices $\{X_i\}_{i=1}^n$ are respectively projected to the predicted lncRNA-protein interaction matrix R by using the projection matrices $\{G_i\}_{i=1}^n$. We estimate the projection matrices $\{G_i\}_{i=1}^n$ for features by minimizing the squared error between their products and the predicted lncRNA-protein interaction matrix R. So we have:

$$\min_{G_i} \sum_{i=1}^{n} \|X_i G_i^T - R\|_F^2 \quad \text{s.t.} G_i \geq 0 \tag{3}$$

where $\|\cdot\|_F^2$ is the Frobenius norm, and the projection matrices $\{G_i\}_{i=1}^n$ are required to be nonnegative.

Then, we introduce the $\ell_{1,2}$-norm regularization term of $\{G_i\}_{i=1}^n$ to ensure the smoothness of the projection matrices. The predicted matrix R should be approximated to the known interaction matrix Y. We can have

$$\min_{G_i,R} \|R-Y\|_F^2 + \mu\sum_{i=1}^n \|X_i G_i^T - R\|_F^2 + \lambda\sum_{i=1}^n \|G_i\|_{1,2}^2 \quad \text{s.t.} G_i \geq 0 \tag{4}$$

where λ is the regularization coefficient, and μ is a trade-off parameter. $\|G_i\|_{1,2} = \sqrt{\sum_k (\sum_l |g_{k,l}^i|)^2}$.

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:

$$\sum_{i=1}^m \theta_i^\eta tr(R^T(D_i - W_i)R) \tag{5}$$

where D_i is a diagonal matrix whose diagonal elements are corresponding row sums of W_i, and $\theta = [\theta_1,\theta_2,\cdots,\theta_i,\cdots,\theta_m]$ 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 θ, 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:

$$\min_{G_i,R,\theta} \|R-Y\|_F^2 + \mu\sum_{i=1}^n \|X_i G_i^T - R\|_F^2 + \sum_{i=1}^m \theta_i^\eta tr(R^T(D_i - W_i)R) + \lambda\sum_{i=1}^n \|G_i\|_{1,2}^2 \quad \text{s.t.} G_i \geq 0, \sum_i \theta_i = 1 \tag{6}$$

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

$$\begin{aligned} Lf = {} & \|R-Y\|_F^2 + \mu\textstyle\sum_{i=1}^n \|X_i G_i^T - R\|_F^2 + \sum_{i=1}^m \theta_i^\eta tr(R^T(D_i - W_i)R) + \lambda\sum_{i=1}^n \|G_i\|_{1,2}^2 \\ & - \delta(\textstyle\sum_{i=1}^m \theta_i - 1) - \sum_{i=1}^n tr(\Gamma_i G_i) \end{aligned}$$

We calculate the partial derivatives of above function with respect to R, G_i and θ_i, and obtain the update rules about R, θ_i and G_i (proof and deduction are provided in S1 File):

$$R = (\textstyle\sum_{i=1}^m \theta_i^\eta (D_i - W) + (1+n\mu)I)^{-1}(Y + \mu\sum_{i=1}^n X_i G_i^T) \tag{7}$$

$$\theta_i = \frac{\left(\frac{1}{tr(R^T(D_i - W_i)R)}\right)^{\frac{1}{\eta-1}}}{\sum_i^m \left(\frac{1}{tr(R^T(D_i - W_i)R)}\right)^{\frac{1}{\eta-1}}} \tag{8}$$

$$G_i = G_i \bigodot \sqrt{\frac{G_i(\mu X_i^T X_i + \lambda ee^T)^+ + \mu(R^T X_i)^-}{G_i(\mu X_i^T X_i + \lambda ee^T)^- + \mu(R^T X_i)^+}} \tag{9}$$

where e is a column vector with all elements equal to 1, and has the same column dimensions

as X_i. $\odot$ 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

$$A^+ = \frac{(|A| + A)}{2}, A^- = \frac{(|A| - A)}{2} \tag{10}$$

Thus, we update R, G_i and θ_i 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}, θ_{lu} and θ_{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, $\{X_{l1}, X_{l2}, \ldots, X_{ln}\}$; protein feature matrices, $\{X_{p1}, X_{p2}, \ldots, X_{pn}\}$; lncRNA normalized similarity matrices, $\{W_{l1}, W_{l2}, \ldots, W_{lm}\}$; protein normalized similarity matrices, $\{W_{p1}, W_{p2}, \ldots, W_{pm}\}$; 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 $\{G_{l1}, G_{l2}, \ldots, G_{ln}\}$; projection matrices of protein features $\{G_{p1}, G_{p2}, \ldots, G_{pn}\}$; weighting parameters of lncRNA similarity matrices, $\{\theta_{l1}, \theta_{l2}, \ldots, \theta_{lm}\}$; weighting parameters of protein similarity matrices, $\{\theta_{p1}, \theta_{p2}, \ldots, \theta_{pm}\}$;

Initialize:

for each $i(1 \le i \le n)$
  initialize G_{li}, G_{pi} with random values on interval [0,1];
end for
for each $i(1 \le i \le m)$
  initialize θ_{li}, θ_{pi} as $1/m$;
end for
repeat
  update R_l via (7) with fixing $\{G_{li}\}_{i=1}^n$, $\{\theta_{li}\}_{i=1}^m$;
  for each $i(1 \le i \le n)$
    update G_{li} via (8) with fixing R_l;
  end for
  for each $i(1 \le i \le m)$
    update θ_{li} via (9) with fixing R_l:
  end for
until Converges;
repeat

update R_p via (7) with fixing $\{G_{pi}\}_{i=1}^{n}$, $\{\theta_{pi}\}_{i=1}^{m}$;
for each $i(1 \le i \le n)$
update G_{pi} via (8) with fixing R_p;
end for
for each $i(1 \le i \le m)$
update θ_{pi} via (9) with fixing R_p:
end for
until Converges;
$M = (R_l + (R_p)^T)/2$
Return M

Algorithm 2: Predicting interacting proteins (or lncRNAs) for new lncRNAs (or proteins) by SFPEL-LPI

Input: feature matrices for new lncRNAs, $\{X_{l1}, X_{l2}, \ldots, X_{lu}\}$ (or feature matrices for new proteins, $\{X_{p1}, X_{p2}, \ldots, X_{pv}\}$); projection matrices of lncRNA features $\{G_{l1}, G_{l2}, \ldots, G_{lu}\}$ (or projection matrices of protein features $\{G_{p1}, G_{p2}, \ldots, G_{pv}\}$); weighting parameters of lncRNA similarity matrices, $\{\theta_{l1}, \theta_{l2}, \ldots, \theta_{lu}\}$ (or weighting parameters of protein features, $\{\theta_{p1}, \theta_{p2}, \ldots, \theta_{pv}\}$); ($\{G_{li}\}_{i=1}^{u}$, $\{\theta_{li}\}_{i=1}^{u}$ or $\{G_{pi}\}_{i=1}^{v}$, $\{\theta_{pi}\}_{i=1}^{v}$ are obtained by Algorithm 1);

Output: predicted lncRNA-protein interaction matrix, $R_l = \sum_{i=1}^{u} \theta_{li} X_{li} G_{li}^{T}$ (or predicted protein-lncRNA interaction matrix, $R_p = \sum_{i=1}^{v} \theta_{pi} X_{pi} G_{pi}^{T}$);

Results

Evaluation metrics

We adopt five-fold cross validation to evaluate the performances of prediction models. The proposed method SFPEL-LPI can predict unobserved interactions between known lncRNAs and known proteins, and also can make predictions for new lncRNAs (or proteins). In predicting unobserved lncRNA-protein interactions, all known lncRNA-protein interactions are randomly split into five subsets with equal size. Each time, four subsets are combined as training set and the remaining one subset is used as the testing set. In predicting proteins interacting with new lncRNAs, all known lncRNAs are split into five subsets with equal size. The model is constructed based on the lncRNAs in training set and their interactions with all proteins, and then is used to predict proteins interacting with testing lncRNAs. Similarly, we evaluate the performances of models in predicting lncRNAs interacting with new proteins. Hence, we introduce notations for above mentioned cross validation settings. CV_{lp}: known lncRNA-protein interactions are split into five folds in predicting unobserved interactions. CV_l: known lncRNAs are split into five folds in predicting interactions for new lncRNAs. CV_p: known proteins are split into five folds in predicting interactions for new proteins.

The area under ROC curve (AUC) and the area under precision-recall curve (AUPR) are popular metrics for evaluating prediction models. Since known lncRNA-protein interactions are much less than non-interacting lncRNA-protein pairs, we adopt AUPR as the primary metric, which punishes false positive more in the evaluation process[68,69]. Moreover, we adopt several binary classification metrics, i.e. recall (REC), accuracy (ACC), precision (PR) and F1-measure (F1).

Parameter setting

SFPEL-LPI has three parameters: μ, λ and η. μ is a parameter for the error between projected interactions and predicted lncRNA-protein interactions; λ controls the contribution of projection matrix; η describes strength of different similarity measures.

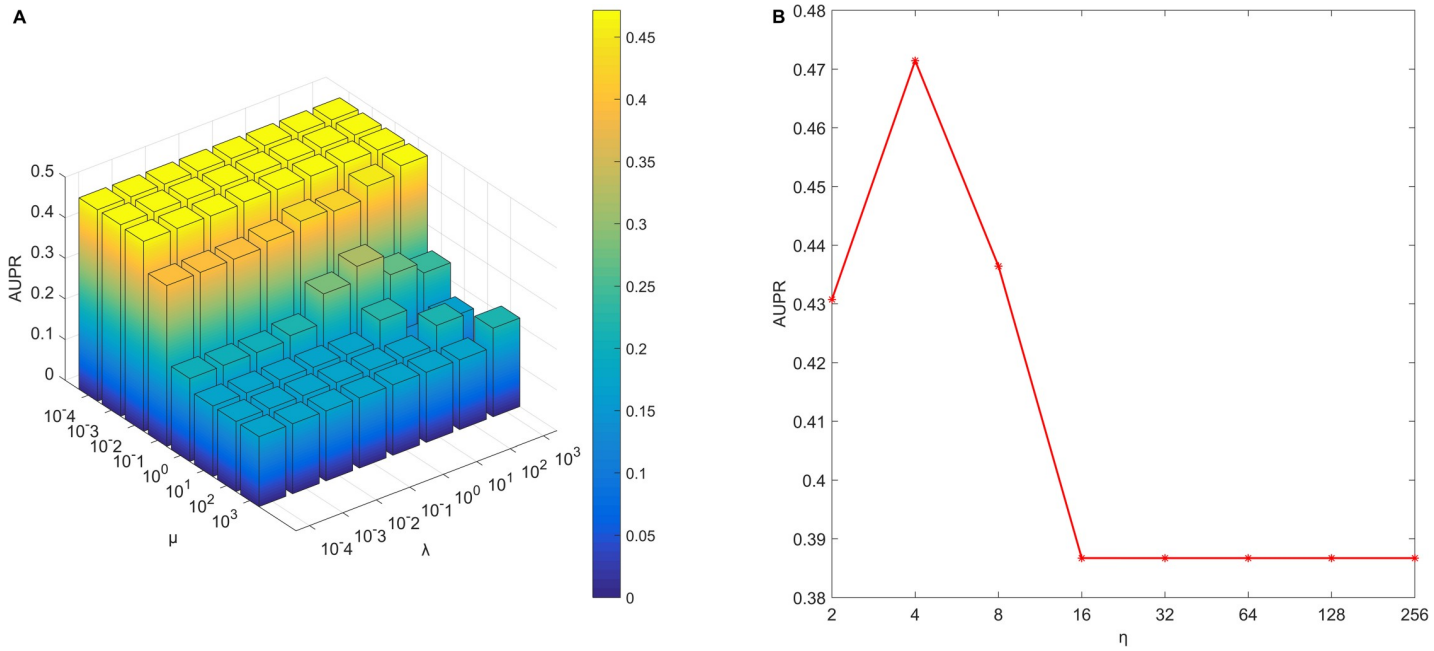

Fig 2. The influence of parameters on AUPR of models. (A) Fix the parameter $\eta = 2^2$, and evaluate the influence of parameters μ and λ. (B) Fix the parameter $\mu = 10^{-1}$, $\lambda = 10^3$, and evaluate the influence of parameter η.

https://doi.org/10.1371/journal.pcbi.1006616.g002

To test influence of parameters, we consider all combinations of parameters $\mu \in \{10^{-4},10^{-3},10^{-2},10^{-1},10^{0},10^{1},10^{2},10^{3}\}$, $\mu \in \{10^{-4},10^{-3},10^{-2},10^{-1},10^{0},10^{1},10^{2},10^{3}\}$ and $\eta \in \{2^1,2^2,2^3,2^4,2^5,2^6,2^7,2^8\}$. We build SFPEL-LPI models by using different parameters, and implement five-fold cross validation CV_{lp} to evaluate SFPEL-LPI models. SFPEL-LPI produces the best AUPR score of 0.473 when $\mu = 10^{-3}$, $\lambda = 10^{-4}$ and $\eta = 2^2$. Then, we fix the parameter $\eta = 2^2$, and evaluate the influence of μ and λ. As shown in Fig 2A, μ greatly influences the performance of SFPEL-LPI, and a smaller value for μ is likely to produce better result. Further, we fix the parameters $\mu = 10^{-3}$ and $\lambda = 10^{-4}$ and test the influence of η. As illustrated in Fig 2B, the performances of SFPEL-LPI decrease as η increases, and then remain unchanged after a threshold.

The parameter η 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 η and lncRNA similarity measures θ_{lncRNA} (or protein similarity measures $\theta_{protein}$). 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 η, 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.

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

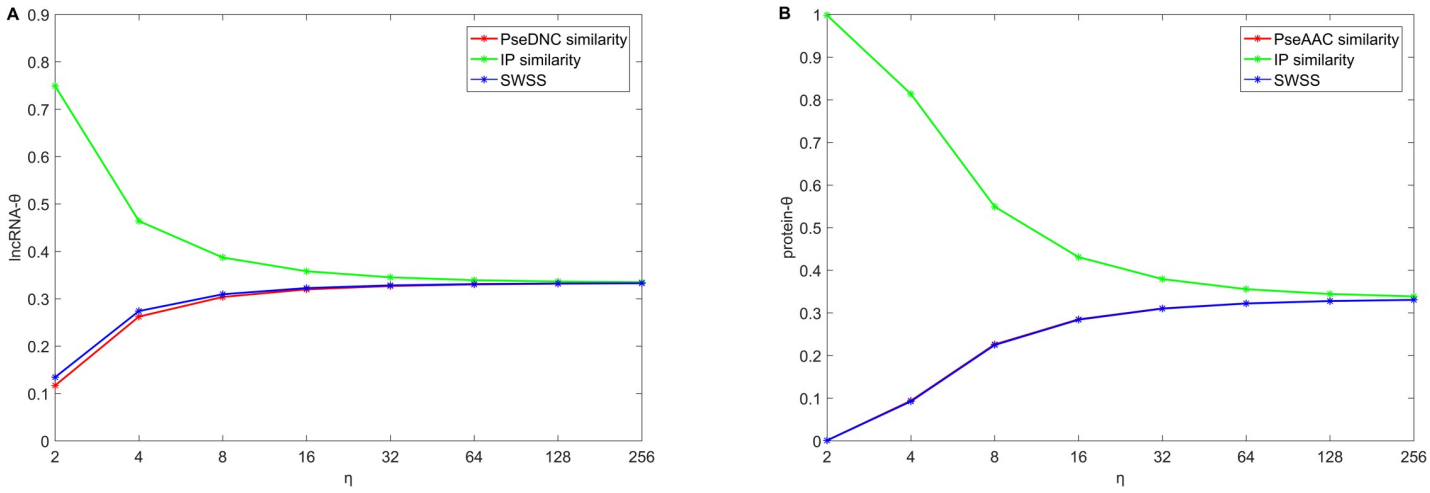

Fig 3. (A) The relationship between η and θ_{lncRNA}. (B) The relationship between η and $\theta_{protein}$.

https://doi.org/10.1371/journal.pcbi.1006616.g003

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 matric "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.

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

Table 1. Performances of SFPEL-LPI for predicting lncRNA-protein interactions.

Cross Validation	AUPR	AUC	PRE	REC	ACC	F1
CV_{lp}	0.473	0.920	0.449	0.495	0.960	0.470
CV_l	0.490	0.823	0.449	0.552	0.823	0.493
CV_p	0.339	0.656	0.325	0.476	0.749	0.375

https://doi.org/10.1371/journal.pcbi.1006616.t001

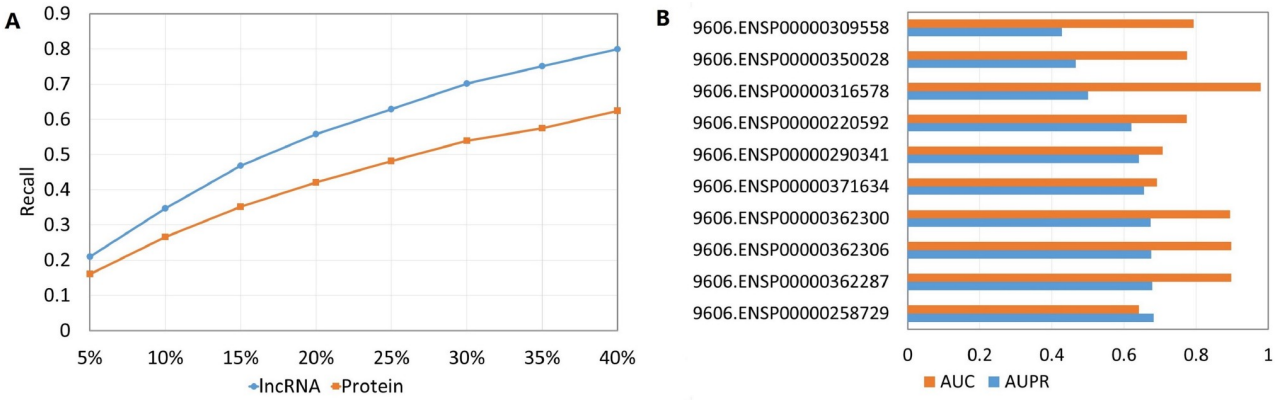

Fig 4. (A) The average recalls in predicting new lncRNAs (or proteins) at different top-ranked percentiles under CV_l or CV_p. (B) The AUC value and AUPR value of predicting interacting lncRNAs for selected new proteins.

https://doi.org/10.1371/journal.pcbi.1006616.g004

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. LPNLP 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_{lp}. 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.

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

Table 2. Performances of prediction methods on the benchmark dataset.

Method	AUPR	AUC	PRE	REC	ACC	F1
RWR	0.236	0.850	0.245	0.391	0.935	0.299
LPBNI	0.330	0.856	0.413	0.370	0.958	0.386
KATZLGO	0.286	0.760	0.354	0.348	0.954	0.350
LPI-ETSLP	0.322	0.889	0.374	0.423	0.953	0.394
LPLNP	0.459	0.910	0.523	0.404	0.965	0.453
SFPEL-LPI	0.473	0.920	0.449	0.495	0.960	0.470

https://doi.org/10.1371/journal.pcbi.1006616.t002

Table 3. Difference between SFPEL-LPI and benchmark methods tested by Paired t-test in terms of AUPR and AUC.

AUPR				
RWR	LPBNI	KATZLGO	LPI-ETSLP	LPLNP
6.35E-37	3.55E-32	1.91E-34	3.37E-31	4.38E-12
		AUC		
RWR	LPBNI	KATZLGO	LPI-ETSLP	LPLNP
1.43E-26	5.94E-28	8.15E-34	1.59E-31	1.37E-19

https://doi.org/10.1371/journal.pcbi.1006616.t003

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 the 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 lncRNA-protein interactions to test the robustness of prediction methods. To be more specific, we randomly remove 5% of known lncRNA-protein interactions and add the same number of inexistent 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.

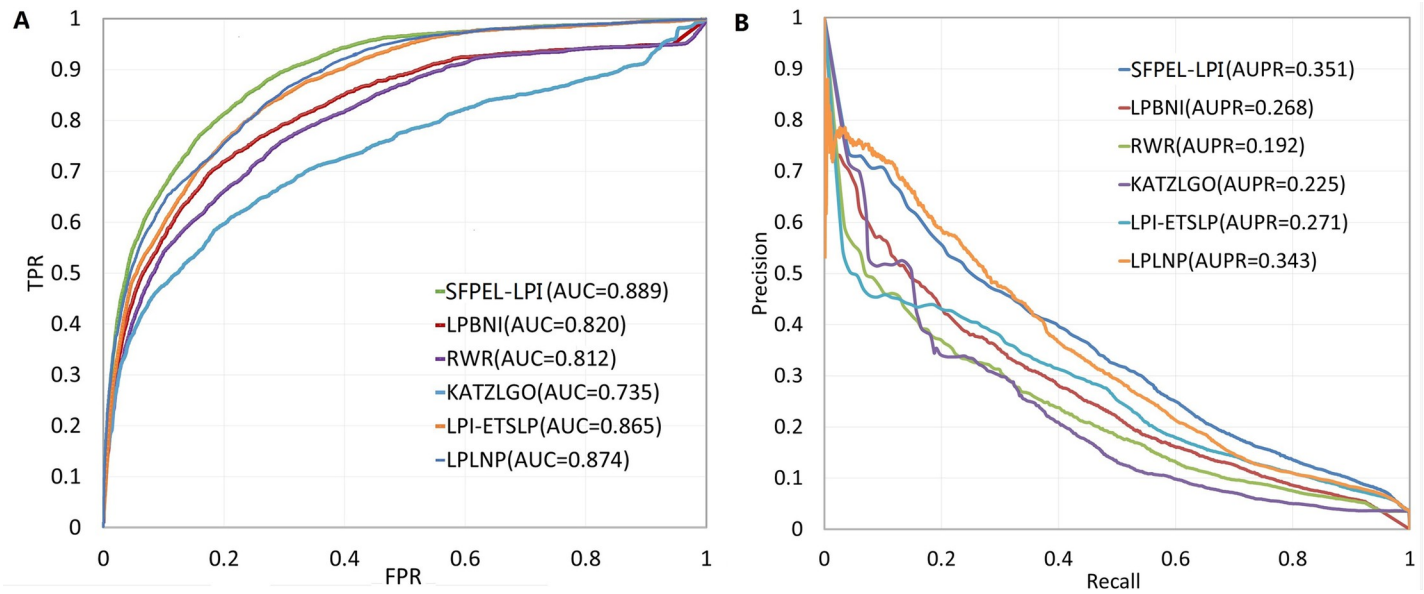

Fig 5. Performance of different methods on the perturbed dataset. (A) ROC curves. (B) PR curves.

https://doi.org/10.1371/journal.pcbi.1006616.g005

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 lncRNA-protein interactions. Therefore, we train the prediction model based on the NPInter v2.0 dataset and predict new lncRNA-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

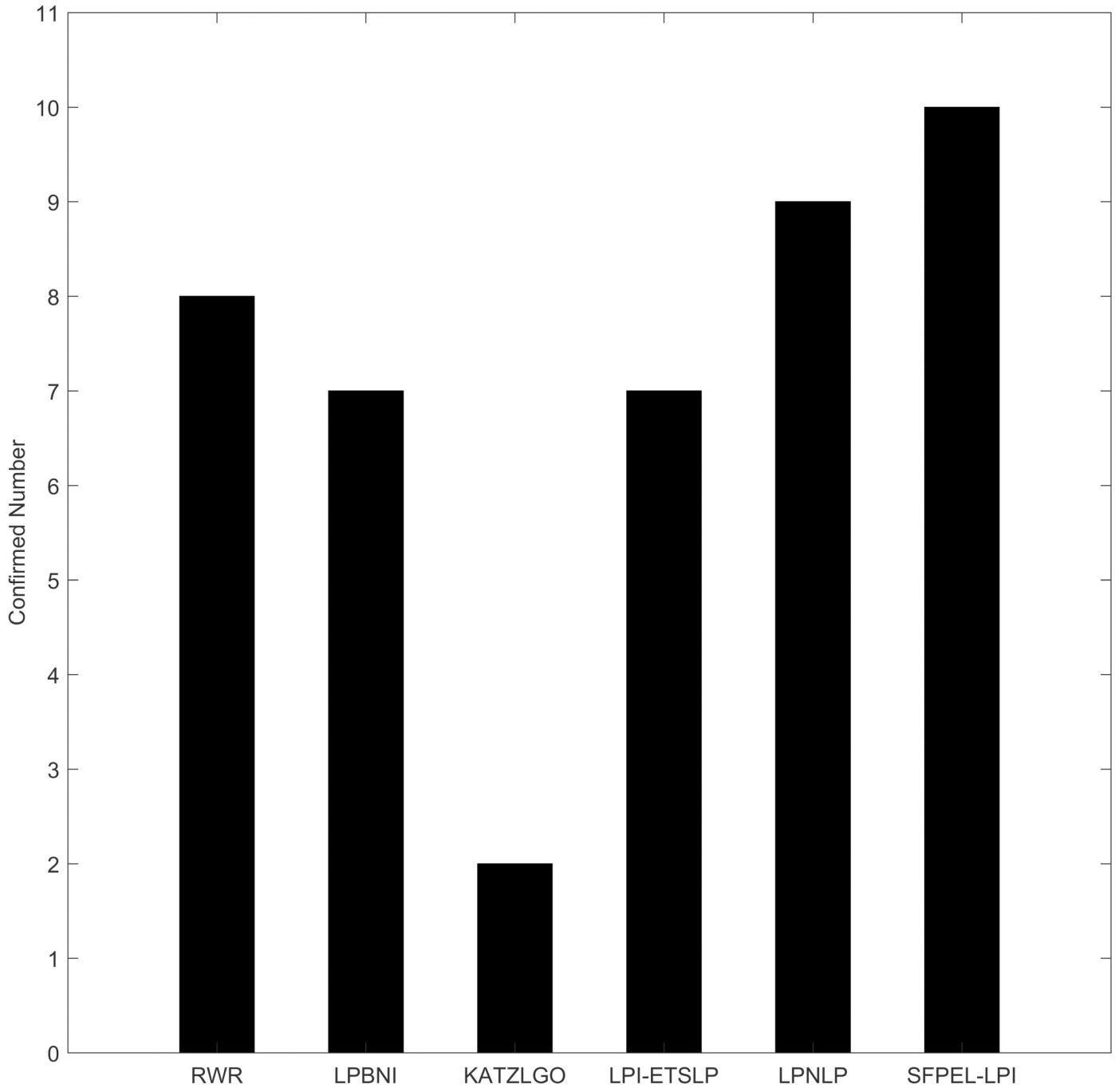

Fig 6. The number of confirmed lncRNA-protein interactions in top 20 predictions of different methods.

https://doi.org/10.1371/journal.pcbi.1006616.g006

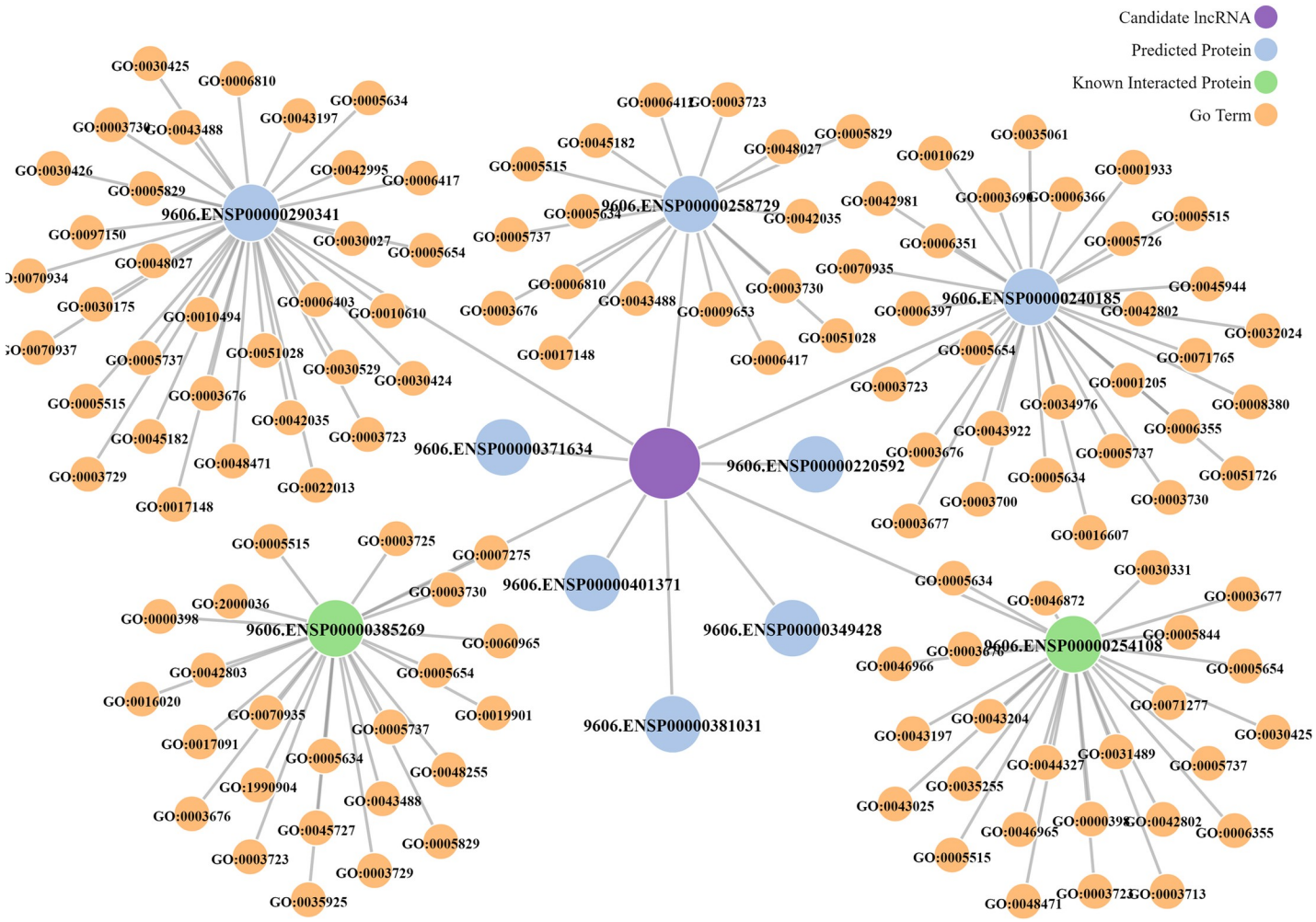

Fig 7. Visualization of top 10 predicted interacting proteins for the lncRNA: "NONHSAT041930". Purple node stands for the lncRNA. Navy blue nodes indicate the predicted interacting proteins, and green nodes represent proteins that have observed interactions with the lncRNA. Moreover, we map the corresponding GO Terms (Orange nodes) of each interacting protein from QuickGO database (https://www.ebi.ac.uk/QuickGO/).

https://doi.org/10.1371/journal.pcbi.1006616.g007

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 lncRNA-protein interaction prediction, available at http://www.bioinfotech.cn/SFPEL-LPI/. Users can input lncRNA sequences (or protein sequences) or upload a text file with FASTA-formatted lncRNA 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, IGF2BP1, IGF2BP2, IGF2BP3, EWSR1 and TIA1 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

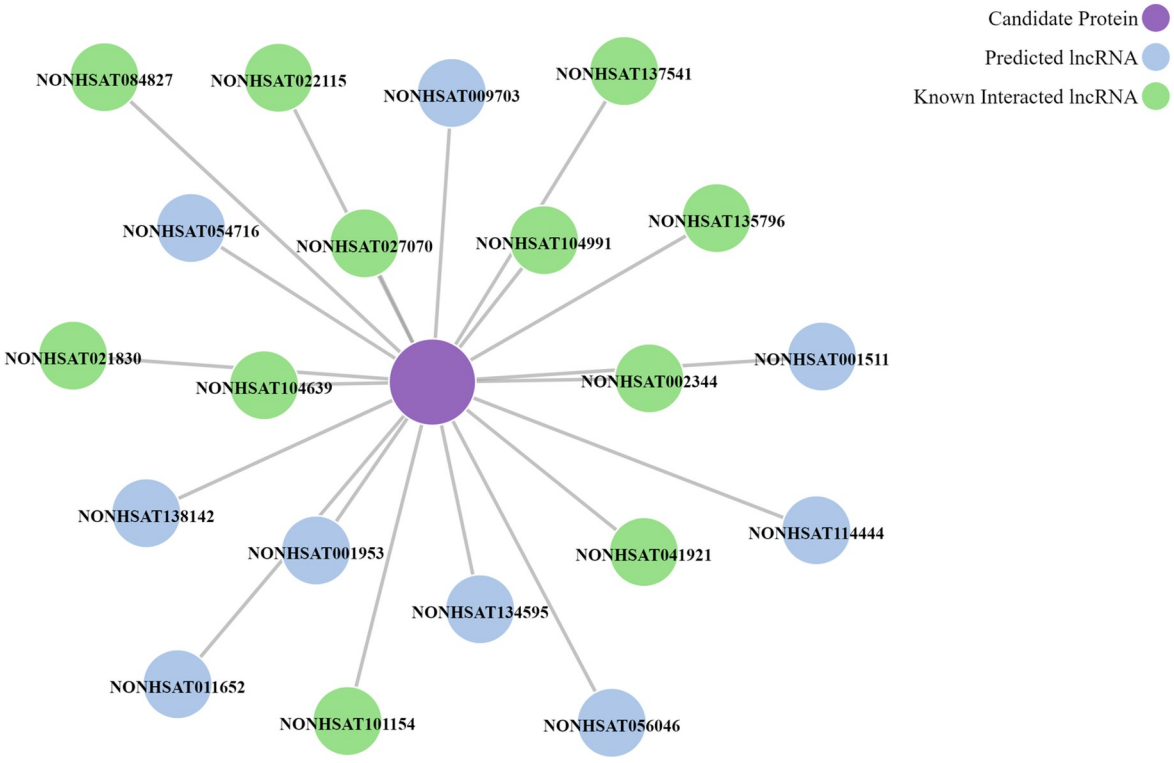

Fig 8. Visualization of top 20 predicted interacting lncRNAs of the protein: 9606.ENSP00000240185. Purple node stands for the protein. Navy blue nodes indicate the predicted interacting lncRNAs and green nodes represent lncRNAs that have observed interactions with the protein.

https://doi.org/10.1371/journal.pcbi.1006616.g008

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_{lp}, CV_l and CV_p, compared with state-of-the-art lncRNA-protein interaction prediction methods. The promising results validate efficacy of the proposed algorithm for predicting lncRNA-protein interactions, especially for the new lncRNAs 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 lncRNA-protein interactions to test the robustness of prediction methods. A web server is constructed to predict interacting proteins/lncRNAs for given lncRNAs/proteins. We adopt the lncRNA "NONHSAT041930" as an example to predict interacting proteins, and can find evidences to confirm novel lncRNA-protein interactions.

However, SFPEL-LPI still has several limitations. It has three parameters, and parameter tuning is time-consuming. In addition, known lncRNA-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.
(PDF)

S2 File. The data of SFPEL-LPI.
(MAT)

S1 Table. Top 20 predictions of SFPEL-LPI and their ranks in predictions of benchmark methods.
(DOCX)

S2 Table. Top 10 interacting proteins of LncRNA "NONHSAT041930" (OIP5-AS1) predicted by SFPEL-LPI.
(DOCX)

S3 Table. Top 20 interacting lncRNAs of protein "9606.ENSP00000240185" (TAR DNA-binding protein 43) predicted by SFPEL-LPI.
(DOCX)

Author Contributions

Conceptualization: Xining Zhang.

Formal analysis: Guifeng Tang, Feng Huang.

Investigation: Wen Zhang, Xiang Yue.

Methodology: Wen Zhang.

Resources: Guifeng Tang.

Software: Wenjian Wu.

Validation: Guifeng Tang.

Visualization: Xiang Yue, Feng Huang.

Writing – original draft: Wen Zhang, Xiang Yue.

References

1. Prensner JR, Chinnaiyan AM (2011) The emergence of lncRNAs in cancer biology. Cancer Discov 1: 391–407. https://doi.org/10.1158/2159-8290.CD-11-0209 PMID: 22096659
2. Volders PJ, Helsens K, Wang X, Menten B, Martens L, et al. (2013) LNCipedia: a database for annotated human lncRNA transcript sequences and structures. Nucleic Acids Res 41: D246–251. https://doi.org/10.1093/nar/gks915 PMID: 23042674
3. Kung JT, Colognori D, Lee JT (2013) Long noncoding RNAs: past, present, and future. Genetics 193: 651–669. https://doi.org/10.1534/genetics.112.146704 PMID: 23463798
4. Pang KC, Frith MC, Mattick JS (2006) Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. Trends Genet 22: 1–5. https://doi.org/10.1016/j.tig.2005.10.003 PMID: 16290135
5. Kutter C, Watt S, Stefflova K, Wilson MD, Goncalves A, et al. (2012) Rapid turnover of long noncoding RNAs and the evolution of gene expression. PLoS Genet 8: e1002841. https://doi.org/10.1371/journal.pgen.1002841 PMID: 22844254
6. Fu M, Zou C, Pan L, Liang W, Qian H, et al. (2016) Long noncoding RNAs in digestive system cancers: Functional roles, molecular mechanisms, and clinical implications (Review). Oncol Rep 36: 1207–1218. https://doi.org/10.3892/or.2016.4929 PMID: 27431376
7. St Laurent G 3rd, Wahlestedt C (2007) Noncoding RNAs: couplers of analog and digital information in nervous system function? Trends Neurosci 30: 612–621. https://doi.org/10.1016/j.tins.2007.10.002 PMID: 17996312
8. Qu Z, Adelson DL (2012) Evolutionary conservation and functional roles of ncRNA. Front Genet 3: 205. https://doi.org/10.3389/fgene.2012.00205 PMID: 23087702
9. Wang KC, Chang HY (2011) Molecular mechanisms of long noncoding RNAs. Mol Cell 43: 904–914. https://doi.org/10.1016/j.molcel.2011.08.018 PMID: 21925379
10. Muppirala UK, Honavar VG, Dobbs D (2011) Predicting RNA-protein interactions using only sequence information. BMC Bioinformatics 12: 489. https://doi.org/10.1186/1471-2105-12-489 PMID: 22192482
11. Wang Y, Chen X, Liu ZP, Huang Q, Wang Y, et al. (2013) De novo prediction of RNA-protein interactions from sequence information. Mol Biosyst 9: 133–142. https://doi.org/10.1039/c2mb25292a PMID: 23138266
12. Suresh V, Liu L, Adjeroh D, Zhou X (2015) RPI-Pred: predicting ncRNA-protein interaction using sequence and structural information. Nucleic Acids Res 43: 1370–1379. https://doi.org/10.1093/nar/gkv020 PMID: 25609700
13. Xiao Y, Zhang J, Deng L (2017) Prediction of lncRNA-protein interactions using HeteSim scores based on heterogeneous networks. Sci Rep 7: 3664. https://doi.org/10.1038/s41598-017-03986-1 PMID: 28623317
14. Lu Q, Ren S, Lu M, Zhang Y, Zhu D, et al. (2013) Computational prediction of associations between long non-coding RNAs and proteins. BMC Genomics 14: 651. https://doi.org/10.1186/1471-2164-14-651 PMID: 24063787
15. Li A, Ge M, Zhang Y, Peng C, Wang M (2015) Predicting Long Noncoding RNA and Protein Interactions Using Heterogeneous Network Model. Biomed Res Int 2015: 671950. https://doi.org/10.1155/2015/671950 PMID: 26839884
16. Yang JH, Li A, Ge MQ, Wang MH (2015) Prediction of interactions between lncRNA and protein by using relevance search in a heterogeneous lncRNA-protein network. 2015 34th Chinese Control Conference (Ccc): 8540–8544.
17. Wiggins BS, Saseen JJ, Page RL 2nd, Reed BN, Sneed K, et al. (2016) Recommendations for Management of Clinically Significant Drug-Drug Interactions With Statins and Select Agents Used in Patients With Cardiovascular Disease: A Scientific Statement From the American Heart Association. Circulation 134: e468–e495. https://doi.org/10.1161/CIR.0000000000000456 PMID: 27754879
18. Zheng XX, Tian K, Wang Y, Guan JH, Zhou SG (2016) Predicting lncRNA-Protein Interactions Based on Protein-Protein Similarity Network Fusion. Bioinformatics Research and Applications, Isbra 2016 9683: 321–322.

19. Zhang Z, Zhang J, Fan C, Tang Y, Deng L (2017) KATZLGO: Large-scale Prediction of LncRNA Functions by Using the KATZ Measure Based on Multiple Networks. IEEE/ACM Trans Comput Biol Bioinform.
20. Hu H, Zhu C, Ai H, Zhang L, Zhao J, et al. (2017) LPI-ETSLP: lncRNA-protein interaction prediction using eigenvalue transformation-based semi-supervised link prediction. Mol Biosyst.
21. Zhang W, Qu Q, Zhang Y, Wang W (2018) The linear neighborhood propagation method for predicting long non-coding RNA–protein interactions. Neurocomputing 273: 526–534.
22. Wei L, Tang J, Zou Q (2017) Local-DPP: An improved DNA-binding protein prediction method by exploring local evolutionary information. Information Sciences 384: 135–144.
23. Song L, Li D, Zeng X, Wu Y, Guo L, et al. (2014) nDNA-prot: identification of DNA-binding proteins based on unbalanced classification. BMC Bioinformatics 15: 298. https://doi.org/10.1186/1471-2105-15-298 PMID: 25196432
24. Yuan J, Wu W, Xie C, Zhao G, Zhao Y, et al. (2014) NPInter v2.0: an updated database of ncRNA interactions. Nucleic Acids Res 42: D104–108. https://doi.org/10.1093/nar/gkt1057 PMID: 24217916
25. Gough J, Karplus K, Hughey R, Chothia C (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. Journal of Molecular Biology 313: 903–919. https://doi.org/10.1006/jmbi.2001.5080 PMID: 11697912
26. Zheng XX, Wang Y, Tian K, Zhou JG, Guan JH, et al. (2017) Fusing multiple protein-protein similarity networks to effectively predict lncRNA-protein interactions. Bmc Bioinformatics 18.
27. Junge A, Refsgaard JC, Garde C, Pan XY, Santos A, et al. (2017) RAIN: RNA-protein Association and Interaction Networks. Database-the Journal Of Biological Databases And Curation.
28. Cheng ZZ, Huang K, Wang Y, Liu H, Guan JH, et al. (2017) Selecting high-quality negative samples for effectively predicting protein-RNA interactions. Bmc Systems Biology 11.
29. Pan XY, Fan YX, Yan JC, Shen HB (2016) IPMiner: hidden ncRNA-protein interaction sequential pattern mining with stacked autoencoder for accurate computational prediction. Bmc Genomics 17.
30. Zhang W, Liu J, Zhao M, Li Q (2012) Predicting linear B-cell epitopes by using sequence-derived structural and physicochemical features. International Journal of Data Mining and Bioinformatics 6: 557–569. PMID: 23155782
31. Zhang W, Niu Y, Zou H, Luo L, Liu Q, et al. (2015) Accurate prediction of immunogenic T-cell epitopes from epitope sequences using the genetic algorithm-based ensemble learning. PLoS One 10: e0128194. https://doi.org/10.1371/journal.pone.0128194 PMID: 26020952
32. Zhang W, Chen Y, Liu F, Luo F, Tian G, et al. (2017) Predicting potential drug-drug interactions by integrating chemical, biological, phenotypic and network data. BMC Bioinformatics 18: 18. https://doi.org/10.1186/s12859-016-1415-9 PMID: 28056782
33. Zhang W, Yue X, Chen Y, Lin W, Li B, et al. Predicting drug-disease associations based on the known association bipartite network; 2017. pp. 503–509.
34. Zhang W, Yue X, Liu F, Chen Y, Tu S, et al. (2017) A unified frame of predicting side effects of drugs by using linear neighborhood similarity. BMC Systems Biology 11: 101. https://doi.org/10.1186/s12918-017-0477-2 PMID: 29297371
35. Li D, Luo L, Zhang W, Liu F, Luo F (2016) A genetic algorithm-based weighted ensemble method for predicting transposon-derived piRNAs. BMC Bioinformatics 17: 329. https://doi.org/10.1186/s12859-016-1206-3 PMID: 27578422
36. Luo L, Li D, Zhang W, Tu S, Zhu X, et al. (2016) Accurate Prediction of Transposon-Derived piRNAs by Integrating Various Sequential and Physicochemical Features. PLOS ONE 11.
37. Liu B, Liu F, Wang X, Chen J, Fang L, et al. (2015) Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. Nucleic Acids Res 43: W65–71. https://doi.org/10.1093/nar/gkv458 PMID: 25958395
38. Liu B (2017) BioSeq-Analysis: a platform for DNA, RNA and protein sequence analysis based on machine learning approaches. Brief Bioinform.
39. Liu B, Liu F, Fang L, Wang X, Chou KC (2016) repRNA: a web server for generating various feature vectors of RNA sequences. Mol Genet Genomics 291: 473–481. https://doi.org/10.1007/s00438-015-1078-7 PMID: 26085220
40. Liu YP, Wu HY, Yang X, Xu HQ, Li YC, et al. (2015) Association between thiopurine S-methyltransferase polymorphisms and thiopurine-induced adverse drug reactions in patients with inflammatory bowel disease: a meta-analysis. PLoS One 10: e0121745. https://doi.org/10.1371/journal.pone.0121745 PMID: 25799415

41. Liu B, Fang L, Liu F, Wang X, Chou KC (2016) iMiRNA-PseDPC: microRNA precursor identification with a pseudo distance-pair composition approach. J Biomol Struct Dyn 34: 223–235. https://doi.org/10.1080/07391102.2015.1014422 PMID: 25645238

42. Du PF, Zhao W, Miao YY, Wei LY, Wang L (2017) UltraPse: A Universal and Extensible Software Platform for Representing Biological Sequences. Int J Mol Sci 18.

43. Chen W, Feng PM, Lin H, Chou KC (2013) iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. Nucleic Acids Res 41: e68. https://doi.org/10.1093/nar/gks1450 PMID: 23303794

44. Chen W, Feng PM, Lin H, Chou KC (2014) iSS-PseDNC: identifying splicing sites using pseudo dinucleotide composition. Biomed Res Int 2014: 623149. https://doi.org/10.1155/2014/623149 PMID: 24967386

45. Dong C, Yuan YZ, Zhang FZ, Hua HL, Ye YN, et al. (2016) Combining pseudo dinucleotide composition with the Z curve method to improve the accuracy of predicting DNA elements: a case study in recombination spots. Mol Biosyst 12: 2893–2900. https://doi.org/10.1039/c6mb00374e PMID: 27410247

46. Zhang W, Shi J, Tang G, Wu W, Yue X, et al. Predicting small RNAs in bacteria via sequence learning ensemble method; 2017. IEEE. pp. 643–647.

47. Chou KC (2001) Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins 43: 246–255. PMID: 11288174

48. Lin H, Ding H (2011) Predicting ion channels and their types by the dipeptide mode of pseudo amino acid composition. J Theor Biol 269: 64–69. https://doi.org/10.1016/j.jtbi.2010.10.019 PMID: 20969879

49. Liu B, Wang S, Wang X (2015) DNA binding protein identification by combining pseudo amino acid composition and profile-based protein representation. Sci Rep 5: 15479. https://doi.org/10.1038/srep15479 PMID: 26482832

50. Zhang W, Yue X, Huang F, Liu R, Chen Y, et al. (2018) Predicting drug-disease associations and their therapeutic function based on the drug-disease association bipartite network. Methods 145: 51–59. https://doi.org/10.1016/j.ymeth.2018.06.001 PMID: 29879508

51. Smith TF, Waterman MS, Burks C (1985) The statistical distribution of nucleic acid similarities. Nucleic Acids Res 13: 645–656. PMID: 3871073

52. Zeng J, Li D, Wu Y, Zou Q, Liu X (2016) An empirical study of features fusion techniques for protein-protein interaction prediction. Current Bioinformatics 11: 4–12.

53. Zhang W., Yang W., Lu X., Huang F., and Luo F., "The Bi-Direction Similarity Integration Method for Predicting Microbe-Disease Associations," IEEE Access, vol. 6, pp. 38052–38061, 2018.

54. Zhang W, Chen YL, Tu SK, Liu F, Qu QL (2016) Drug side effect prediction through linear neighborhoods and multiple data source integration. 2016 Ieee International Conference on Bioinformatics and Biomedicine (Bibm): 427–434.

55. Zhang W, Liu F, Luo LQ, Zhang JX (2015) Predicting drug side effects by multi-label learning and ensemble learning. Bmc Bioinformatics 16.

56. Zhang W., Xiong Y., Zhao M., Zou H., Ye X., and Liu J., "Prediction of conformational B-cell epitopes from 3D structures by random forests with a distance-based feature," Bmc Bioinformatics, vol. 12, no. 1, pp. 341, 2011.

57. Zhang W., Niu Y., Xiong Y., Zhao M., Yu R., and Liu J., "Computational Prediction of Conformational B-Cell Epitopes from Antigen Primary Structures by Ensemble Learning," Plos One, vol. 7, no. 8, pp. e43575, 2012.

58. Zhang W, Zou H, Luo LQ, Liu QC, Wu WJ, et al. (2016) Predicting potential side effects of drugs by recommender methods and ensemble learning. Neurocomputing 173: 979–987.

59. Xu X, Shen F, Yang Y, Zhang D, Shen HT, et al. (2017) Matrix Tri-Factorization with Manifold Regularizations for Zero-Shot Learning. pp. 2007–2016.

60. Nie F, Wang X, Deng C, Huang H (2017) Learning A Structured Optimal Bipartite Graph for Co-Clustering. pp. 4132–4141.

61. Bai Z, Walker PB, Tschiffely AE, Wang F, Davidson I (2017) Unsupervised Network Discovery for Brain Imaging Data. pp. 55–64.

62. Cai D, He X, Han J, Huang TS (2011) Graph Regularized Nonnegative Matrix Factorization for Data Representation. IEEE Transactions on Pattern Analysis and Machine Intelligence 33: 1548–1560. https://doi.org/10.1109/TPAMI.2010.231 PMID: 21173440

63. Cai D, He X, Han J (2007) Spectral Regression: A Unified Approach for Sparse Subspace Learning. pp. 73–82.

64. Wang Y, Qi W, Zhang L, Ying Z, Sha O, et al. (2017) The novel targets of DL-3-n-butylphthalide predicted by similarity ensemble approach in combination with molecular docking study. Quant Imaging Med Surg 7: 532–536. https://doi.org/10.21037/qims.2017.10.08 PMID: 29184765

65. Wang Z, Liang L, Yin Z, Lin J (2016) Improving chemical similarity ensemble approach in target prediction. J Cheminform 8: 20. https://doi.org/10.1186/s13321-016-0130-x PMID: 27110288

66. Zhou B, Sun Q, Kong DX (2016) Predicting cancer-relevant proteins using an improved molecular similarity ensemble approach. Oncotarget 7: 32394–32407. https://doi.org/10.18632/oncotarget.8716 PMID: 27083051

67. Chen B, McConnell KJ, Wale N, Wild DJ, Gifford EM (2011) Comparing bioassay response and similarity ensemble approaches to probing protein pharmacology. Bioinformatics 27: 3044–3049. https://doi.org/10.1093/bioinformatics/btr506 PMID: 21903625

68. Davis J, Goadrich M (2006) The relationship between Precision-Recall and ROC curves. Proceedings of the 23rd international conference on Machine learning. Pittsburgh, Pennsylvania, USA: ACM. pp. 233–240.

69. Zhang W, Yue X, Lin W, Wu W, Liu R, et al. (2018) Predicting drug-disease associations by using similarity constrained matrix factorization. BMC bioinformatics 19: 233. https://doi.org/10.1186/s12859-018-2220-4 PMID: 29914348

70. Natarajan N, Dhillon IS (2014) Inductive matrix completion for predicting gene-disease associations. Bioinformatics 30: i60–68. https://doi.org/10.1093/bioinformatics/btu269 PMID: 24932006

71. Singh-Blom UM, Natarajan N, Tewari A, Woods JO, Dhillon IS, et al. (2013) Prediction and validation of gene-disease associations using methods inspired by social network analyses. PLoS One 8: e58977. https://doi.org/10.1371/journal.pone.0058977 PMID: 23650495

72. Luo Y, Zhao X, Zhou J, Yang J, Zhang Y, et al. (2017) A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. Nat Commun 8: 573. https://doi.org/10.1038/s41467-017-00680-8 PMID: 28924171

73. van Heesch S, van Iterson M, Jacobi J, Boymans S, Essers PB, et al. (2014) Extensive localization of long noncoding RNAs to the cytosol and mono- and polyribosomal complexes. Genome Biol 15: R6. https://doi.org/10.1186/gb-2014-15-1-r6 PMID: 24393600

74. Ulitsky I, Shkumatava A, Jan CH, Sive H, Bartel DP (2011) Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. Cell 147: 1537–1550. https://doi.org/10.1016/j.cell.2011.11.055 PMID: 22196729

75. Kim J, Abdelmohsen K, Yang X, De S, Grammatikakis I, et al. (2016) LncRNA OIP5-AS1/cyrano sponges RNA-binding protein HuR. Nucleic Acids Res 44: 2378–2392. https://doi.org/10.1093/nar/gkw017 PMID: 26819413

76. Kim J, Noh JH, Lee SK, Munk R, Sharov A, et al. (2017) LncRNA OIP5-AS1/cyrano suppresses GAK expression to control mitosis. Oncotarget 8: 49409–49420. https://doi.org/10.18632/oncotarget.17219 PMID: 28472763

77. Chen M, Zhao H, Lind SB, Pettersson U (2016) Data on the expression of cellular lncRNAs in human adenovirus infected cells. Data Brief 8: 1263–1279. https://doi.org/10.1016/j.dib.2016.06.053 PMID: 27547808

78. Liu X, Zheng J, Xue Y, Yu H, Gong W, et al. (2018) PIWIL3/OIP5-AS1/miR-367-3p/CEBPA feedback loop regulates the biological behavior of glioma cells. Theranostics 8: 1084–1105. https://doi.org/10.7150/thno.21740 PMID: 29464001