life

MDPI

Article

PRIP: A Protein-RNA Interface Predictor Based on Semantics of Sequences

You Li [1], Jianyi Lyu [1], Yaoqun Wu [2], Yuewu Liu [3] and Guohua Huang [1,*]

[1] School of Electrical Engineering, Shaoyang University, Shaoyang 422000, China; youli9609@163.com (Y.L.); ljy990309@163.com (J.L.)
[2] Hunan Key Laboratory for Computation and Simulation in Science and Engineering, Xiangtan University, Xiangtan 411105, China; 201731510071@smail.xtu.edu.cn
[3] College of Information and Intelligence, Hunan Agricultural University, Changsha 410128, China; yuewuliu@whu.edu.cn
* Correspondence: 3280@hnsyu.edu.cn or guohuahhn@163.com

Abstract: RNA–protein interactions play an indispensable role in many biological processes. Growing evidence has indicated that aberration of the RNA–protein interaction is associated with many serious human diseases. The precise and quick detection of RNA–protein interactions is crucial to finding new functions and to uncovering the mechanism of interactions. Although many methods have been presented to recognize RNA-binding sites, there is much room left for the improvement of predictive accuracy. We present a sequence semantics-based method (called PRIP) for predicting RNA-binding interfaces. The PRIP extracted semantic embedding by pre-training the Word2vec with the corpus. Extreme gradient boosting was employed to train a classifier. The PRIP obtained a SN of 0.73 over the five-fold cross validation and a SN of 0.67 over the independent test, outperforming the state-of-the-art methods. Compared with other methods, this PRIP learned the hidden relations between words in the context. The analysis of the semantics relationship implied that the semantics of some words were specific to RNA-binding interfaces. This method is helpful to explore the mechanism of RNA–protein interactions from a semantics point of view.

Keywords: RNA-protein interactions; word2vec; xgboost; embedding; semantics

Citation: Li, Y.; Lyu, J.; Wu, Y.; Liu, Y.; Huang, G. PRIP: A Protein-RNA Interface Predictor Based on Semantics of Sequences. *Life* **2022**, *12*, 307. https://doi.org/10.3390/life12020307

Academic Editors: Stefano Gianni and Attila Ambrus

Received: 5 January 2022
Accepted: 4 February 2022
Published: 18 February 2022

1. Introduction

Proteins and RNAs are two of the most important biological macromolecules that constitute life. They exert roles in many biological processes, such as protein synthesis, DNA repair, DNA replication, regulation of gene expression, and viral replication, by interacting with each other [1–5]. Increasing evidence shows that aberrant protein-RNA interactions are closely associated with many complex human diseases [5–17], such as Alzheimer's disease [12–14], tumors [15], lung cancer [16], and cardiovascular diseases [17]. Therefore, precisely identifying protein-RNA interfaces not only helps understand the mechanism of protein-RNA interactions and provides insight into the pathological mechanisms related to diseases, but also contributes to drug discovery and development.

It is a challenging task to quickly and accurately identify the RNA-binding interface [18,19]. The current experimental methods are capable of accurately detecting RNA-binding sites but are very costly and time-consuming. On the other hand, the computational methods are able to screen protein-RNA interfaces inexpensively and on a large scale, but their accuracy is discouraging. In spite of this, computational methods can inform experimental methods in finding potential RNA-binding interfaces. Over the past decades, many computational methods have been developed for predicting RNA-binding interfaces or RNA–protein interactions [18–44]. These computational methods fall into the framework of machine learning, which has two major components: features and algorithms. According to the used features, these computational methods are grouped into three categories:

sequence-based, structure-based, and hybrid methods [45]. The sequence-based methods extract informative features directly from primary protein sequences, including widely used position-specific scoring matrices (PSSMs) [40], which are generally computed using PSI-BlAST [46], physicochemical properties of amino acids, and pseudo amino acid composition [47]. Most of the sequence-based methods are easy to understand and compute, but they are insufficient to characterize RNA-binding interfaces. In addition, the computation of PSSMs requires large-scale reference datasets and thus is time-consuming. The structure-based methods extract structural features that are beneficial for improving the prediction of the RNA-binding interface. However, the actual structures of most proteins are not available, and the structures predicted by most computational methods generally contain noise. The hybrid methods inherit the strength of both the sequence-based and the structure-based methods, but they also absorb their shortcomings. Three types of methods also suffer from local interferences [33].

Protein sequences are very similar to sentences in the natural language, where each word has a semantic context. The advance in natural language processing (NLP) makes it easy to seize the semantics of words from the context. For example, the Word2vec [48,49] translates words into embedding, making it easy to measure semantic relationships between words. The NLP techniques have been successfully applied to a wide range of areas, including sentiment analysis, spam detection, machine translation, and question answering [50], over the past decades. The NLP techniques have also been recently utilized in the area of bioinformatics [51–54]. For instance, the long-short term memory (LSTM), a technique of NLP, along with the Word2vec were used for identifying antibacterial peptides in protein sequences [54] and predicting human–virus protein–protein interactions [51]. For more examples, readers can refer to three relevant reviews [55–57]. Inspired by the success of NLP, we presented a protein-RNA interface predictor based on the semantics of protein sequences (called PRIP). The PRIP used the Word2vec to extract the semantic embedding of protein sequences and employed the extreme gradient boosting (XGBoost) to discriminate between RNA-binding interfaces and non-interfaces.

2. Materials and Methods

2.1. Datasets

For a fair comparison with the state-of-the-art methods, we used the same datasets as aPRBind [20]. Namely, the RB198 [22] was used as the training set and the RB111 [40] as the independent set; both were downloaded from http://ailab-projects2.ist.psu.edu/RNABindRPlus/data.html (Accessed on 13 January 2021). The RB198 compiled by Lewis et al. [58] contains unique 198 protein chains. The RB111 is a recently compiled dataset of RNA-binding protein complexes, consisting of 111 protein chains. In both the RB198 and the RB111, the intra-sequence identities are less than 0.3. The chains in the RB111 have less than 0.4 sequence identity with those in the RB198 and the RB44 constructed by Puton et al. [59]. This is a sufficient reason to use the RB111 as the independent set.

A residue with at least one atom closer than 5 Å to any atoms of RNAs was referred to as the interface residue [20]. According to the definition, the RB198 contains 7950 interface residues and 45,710 non-interface residues, and the RB111 has 3305 interface residues and 34,255 non-interface residues.

2.2. Methodology

As shown in Figure 1, the proposed PRIP consisted of five steps: pre-training the Word2vec [48,49], splitting protein sequences, extracting semantic features, training the XGBoost classifier, and distinguishing between binding and non-binding sites. The corpus of protein sequences was first collected to pre-train the Word2vec. Then, the protein sequences were divided into segments of fixed length. Next, segments were mapped into the semantic features by the pre-trained Word2vec. The semantic features of the training set were used to train an XGBoost classifier. Finally, the trained XGBoost classifier

discriminated the binding from non-binding sites, given the inputs of semantic features of segments.

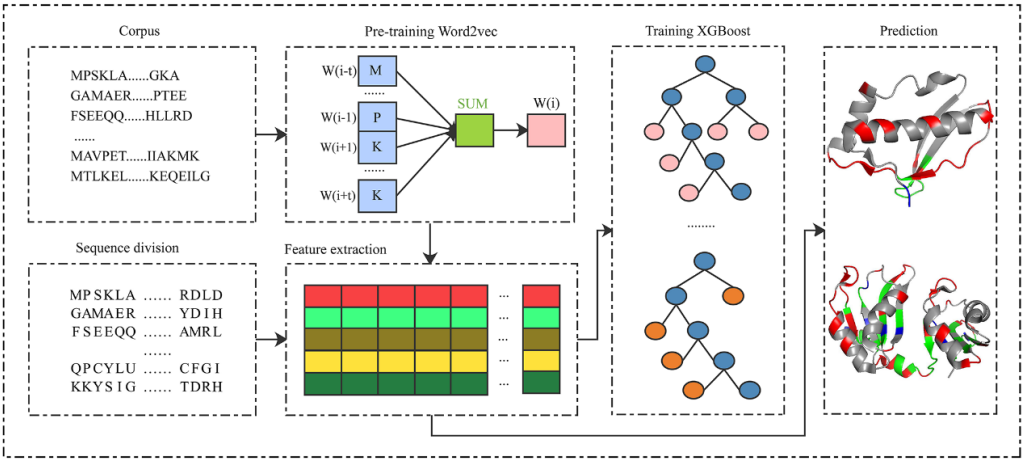

Figure 1. The framework of the PRIP.

2.2.1. Word2vec

Word2vec, proposed by Mikolov et al. [48,49], is a popular algorithm for embedding representations of words. In fact, Word2vec is a general model of neural network that consists of the input, the hidden, and the output layer. The input consists of one-hot encoding vectors, and the theoretical output consists of one-hot representations of words. The input is mapped into the output by multiplying the linking weights between the input and the hidden layer, and then multiplying the linking weights between the hidden and the output layer. The goal of Word2vec is to minimize residues between the theoretical and the actual output. Word2vec has two computational structures: continuous bag-of-word (CBOW) and skip-gram. The CBOW predicts a target word given a context, while the skip-gram does the opposite, namely predict its context given a target word. For each structure, there are two methods of optimization: hierarchical soft-max and negative sampling. When Mikolov et al. [48,49] applied Word2vec to analyze analogical reasoning tasks, some underlying semantic relationships between words were uncovered. An interesting example is that vec("Russia") + vec("river") is close to vec("Volga River"), and vec("Germany") + vec("capital") is close to vec("Berlin"). Due to its efficiency and effectiveness, Word2vec is increasingly attracting attention from the natural language processing community. For more details about Word2vec, readers can refer to the relevant reports [60,61]. Here, we adopted for calculation the Gensim [62], a python tool for the Word2vec algorithm. The Gensim is an open source toolkit available at https://radimrehurek.com/gensim/# (Accessed on 5 March 2021). The parameters of the Word2vec are shown in Table 1.

Table 1. The set of the parameters in Word2vec.

Name	Values
Structure	CBOW
Vector Size	25
Corpus	RB198
Window size	5
Negative sampling	5
Epoch	200
Workers	1

2.2.2. Sequence Division

Protein sequences differ in length. Some are long, while some are short. The protein sequences of variable length are disadvantageous to be subsequently processed by

the machine learning algorithms, because the latter generally requires the input to be length-fixed. Therefore, the primary protein sequences must be divided into length-fixed segments. For each residue in the protein sequence, a length-fixed segment was separated from it. The cut residue was located at the center of the segment, and n residues were located downstream and upstream of the segment. At the start or end of the sequence, the corresponding number of X was added to the segment. For example, a protein sequence was assumed to be TGDFPLO, with n as 3. The protein sequence was split into XXXTGDF, XXTGDFP, XTGDFPL, TGDFPLO, GDFPLOX, DFPLOXX, and FPLOXXX. The segments with the interface residue at the center were positive examples in the training set and the independent set and were otherwise considered negative.

2.2.3. Feature Extraction

We used the 198 protein sequences as the corpus to pre-train Word2vec, where each amino acid was viewed as a word. The pre-trained Word2vec was like a semantic dictionary, where each word (amino acid) corresponded to a semantic vector. Using the semantic dictionary, each residue in the segment was mapped into a semantic vector. Concatenating all the semantic vectors in the segment made it possible to obtain the semantic features of the segment.

2.2.4. XGBoost

XGBoost, proposed by Chen et al. [63], is an improved GBDT (Gradient Boosting Decision tree) algorithm. The XGBoost has the advantages of high efficiency, flexibility, and portability over the traditional GBDT. Similar to the random forest, the XGBoost is an ensemble learning algorithm. The XGBoost generally consists of many decision trees. The outputs of all decision trees are combined as the final output of the XGBoost. Unlike the random forest, the XGBoost is an additive model, where a new decision tree is fitted by the residues between the actual and the sum of all the previous trees.

Given a training set, $\mathrm{D} = \{(x_i, y_i)|x_i \in R^m, y_i \in \mathrm{R}\}$, where n and m denote the number of samples and the dimensions of features, respectively. The XGBoost was assumed to consist of K functions (also called classification or regression trees), namely $\mathrm{F} = \{f^1(x), f^2(x), \ldots, f^K(x)\}$. The predictive output $\hat{y}_i^K$ for the sample x_i is the sum of the output values of all the functions f^k (k = 1, 2, 3, … … , K), namely

$$\hat{y}_i^K = \sum_{k=1}^{K} f^k(x_i), \tag{1}$$

where $f^k(x_i)$ denotes the predictive score of the k-th tree.

Assume that the previous t-1 trees are known. The goal of the XGBoost is to look for the t-th tree so as to minimize the sum of the loss between the predictive and the target output. The objective of the XGBoost is modeled as

$$\mathrm{obj} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^t), \tag{2}$$

where y_i is the target for the sample x_i, and $\hat{y}_i^t$ is the predictive output of all the t trees, which is computed by

$$\hat{y}_i^t = \hat{y}_i^{t-1} + f^t(x_i), \tag{3}$$

The function l denotes the loss function, which measures residues between the predictive output $\hat{y}_i^t$ and the target y_i. In order to reduce or remove over-fitting, regularization is employed. The objective with the regularization is expressed as

$$\begin{aligned} \mathrm{obj} &= \textstyle\sum_{i=1}^{n} l(y_i, \hat{y}_i^t) + \sum_{i=1}^{t} \Omega\left(f^t\right) \\ &= \textstyle\sum_{i=1}^{n} l(y_i, \hat{y}_i^{t-1} + f^t(x_i)) + \Omega\left(f^t\right) + \mathrm{constant}, \end{aligned} \tag{4}$$

where

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 \tag{5}$$

In Equation (5), λ and γ are two user-defined hyper-parameters, T is the number of leaf nodes, and ω_j is the weight of the j-th leaf node. Different from the traditional GBDT, which uses the first-order Taylors, the XGBoost [63] uses the second-order Taylors to approximate the loss function, namely

$$l\left(y_i, \hat{y}_i^{t-1} + f_t(x_i)\right) \approx l\left(y_i, \hat{y}_i^{t-1}\right) + g_i f^t(x_i) + \frac{1}{2} h_i [f^k(x_i)]^2 \tag{6}$$

where g_i is the first-order derivative of the loss function,

$$g_i = \frac{\partial l\left(y_I, \hat{y}_i^{(t-1)}\right)}{\partial \hat{y}_i^{(t-1)}} \tag{7}$$

And h_i is the second-order derivative of the loss function,

$$h_i = \frac{\partial^2 l\left(y_i, \hat{y}_i^{(t-1)}\right)}{\partial \hat{y}_i^{(t-1)} \partial \hat{y}_i^{(t-1)}} \tag{8}$$

Because the constant is not influential on the derivative, the objective is equivalently expressed as

$$\text{obj} = \sum_{i=1}^{n} \left[g_i f^t(x_i) + \frac{1}{2} h_i [f^t f_t^2(x_i)]^2 \right] + \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} \omega_j^2 \tag{9}$$

Let I_j be the set of samples belonging to the j-th leaf node, namely

$$I_j = \{i | q(x_i) = j\} \tag{10}$$

where $q(x_i)$ represents the structure of the t-th decision tree. Let

$$G_j = \sum_{i \in I_j} g_i \tag{11}$$

Be the sum of the first-order derivative over all the samples of the j-th leaf node, and

$$H_j = \sum_{i \in I_j} h_i \tag{12}$$

Be the sum of the second-order derivatives over all the samples of the j-th leaf node. The objective is further simplified as

$$\text{obj} = \sum_{j=1}^{T} \left\{ G_j \omega_j + \frac{1}{2} (\lambda + H_j) \omega_j^2 \right\} + \gamma T \tag{13}$$

Equation (13) is univariate and quadratic. If the structure of the decision tree was fixed, the objective could have the minimum. If, and only if, the weight of the leaf node was set to

$$\omega_j^* = -\frac{G_j}{H_j + \lambda} \tag{14}$$

The minimum of the objective was computed by

$$\text{obj}^* = -\frac{1}{2} \sum_{j=1}^{T} \frac{G_i^2}{H_i + \lambda} + \gamma T. \tag{15}$$

The descriptions above introduced how to optimize the weights of the leaf nodes given the fixed structures of trees. It is easy to understand and realize, but the optimization of the tree structure is an NP-complete question. The number of trees would increase exponentially with the increasing number of samples. It is impossible in practice to

enumerate all possible trees to reach the global optimum solution. A practical solution is to adopt the greedy algorithm. The XGBoost begins with one leaf node and expands to new branches iteratively. The new expanded tree was assumed to be with the left branch L and the right branch R. The gain of the objective was computed by

$$\text{Gain} = \frac{1}{2}\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{1}{2}\frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{1}{2}\frac{\left(\sum_{i\in \mathrm{I}} g_i\right)^2}{\sum_{i\in \mathrm{I}} h_i + \lambda} - \gamma \quad (16)$$

The tree with the minimum gain was chose for the next possible expansion.

2.3. Evaluation Metrics

The k-fold cross validation and the independent test are the commonly used ways of examining the performance of the machine learning algorithms. In the k-fold cross validations, all the training samples are divided into k parts. The machine learning algorithm is trained by k-1 parts of the sample and is tested by the remaining part. The process is repeated k times. The sensitivity (SN), accuracy (ACC), specificity (SP), and Matthews correlation coefficient (MCC) are used to evaluate the performance, which were computed by

$$\text{SN} = \frac{TP}{TP + FN} \quad (17)$$

$$\text{SP} = \frac{TN}{TN + FP} \quad (18)$$

$$\text{ACC} = \frac{TP + TN}{TP + FP + TN + FN} \quad (19)$$

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TN + FN)(TN + FP)(TP + FN)(TP + FP)}} \quad (20)$$

where *TP* and *TN* are the numbers of correctly predicted interfacial residues (binding site) and the numbers of correctly predicted non-interfacial residues (non-binding site), respectively. *FP* and *FN* stand for the number of wrongly predicted interfacial residues and non-interfacial residues, respectively. The receiver operating characteristic (ROC) curve is also employed to visualize performances. The ROC curve draws true positive rates (SN) against false positive rates (1-SP) under various thresholds. The area under ROC curves (AUROC) is used to quantitatively assess the performance.

3. Results

3.1. Parameter Optimization

In order to investigate the impact of the length of the segments on performance, protein sequences were divided into segments ranging from 21 to 39 at an interval of 2. As shown in Figure 2, the AUROC of the segment of 39 amino acid residues is best. Therefore, we set the length of the segment to 39.

3.2. Selection of Models

There are more than 100 machine learning algorithms that have been applied to a wide range of fields. We compared the XGBoost with five popular algorithms: random forest (RF) [64], support vector machine (SVM) [65], logical regression (LR) [66], gradient boosting decision tree (GBDT) [67], and Lightgbm [68]. All the algorithms were trained by the same RB198 and tested by the identical RB111. The ROC curves are shown in Figure 3A. Obviously, the XGBoost is superior to these five algorithms in terms of predicting RNA-binding protein interfaces. We also compared the word embedding of the Word2vec with three common representations: amino acid composition (AAC) [69], dipeptide composition (DPC) [70], and the composition of k-spaced amino acid group pairs (CKSAAPGP) [71]. The AAC calculates the occurring frequency of each amino acid in a given protein or peptide sequence, resulting in a 20-dimensional vector. The DPC calculates the frequency

of amino acid pair occurrence, so it is a 20 × 20 = 400 dimensional vector. The CKSAAPGP computes the frequency of amino acid group pairs separated by K amino acids. Here, K was set to 3, and the five groups were the aliphatic group (G, A, V, L, M, I), aromatic group (F, Y, W), positive charge group (K, R, H), negative charge group (D, E), and uncharged group (S, T, C, P, N, Q) [72], so the dimension of the CKSAAPGP is $5^2 \times 4 = 100$. The ROC curves are shown in Figure 3B. The embedding of the Word2vec is superior to the three representations. Therefore, we chose semantic embedding of the Word2vec as the representations of proteins and XGBoost as the learning algorithm.

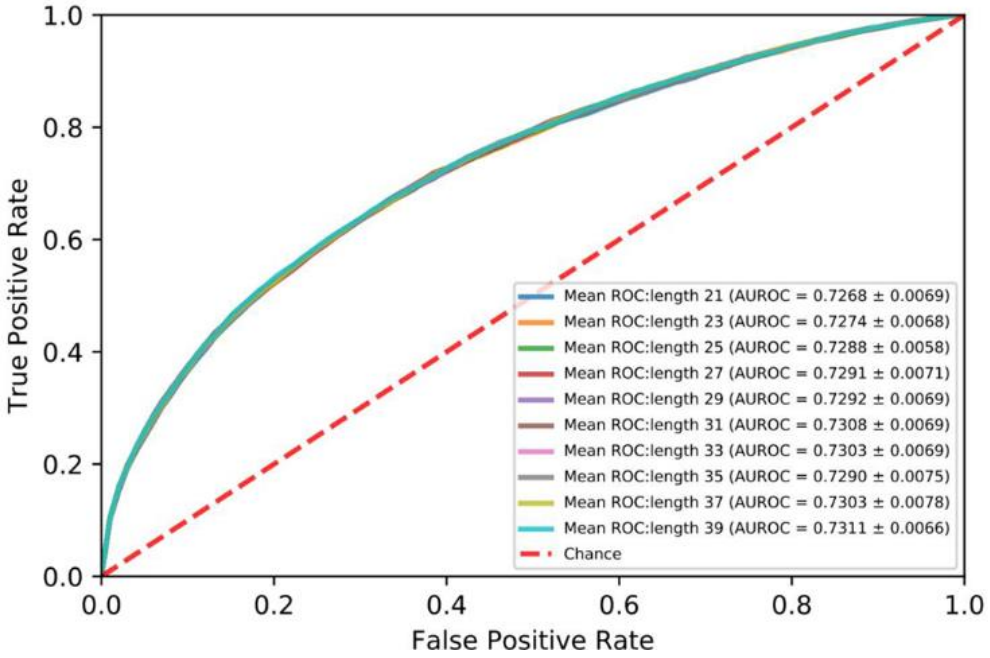

Figure 2. Mean ROC curves for five-fold cross validation: protein sequences 21 to 39.

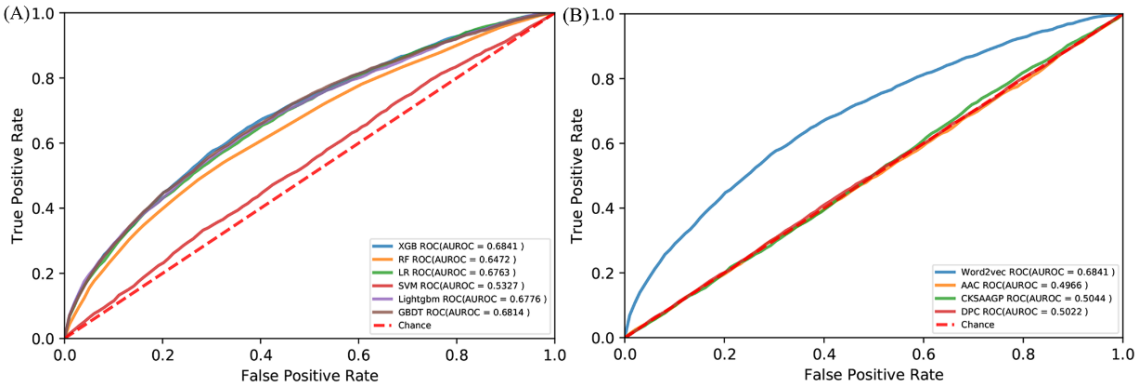

Figure 3. ROC curves of (**A**) different algorithms by the independent test and (**B**) different amino acid encoding methods by the independent test.

3.3. Comparison with State-of-the-Art Methods

The PRIP obtained a mean AUROC of 0.73 over five-fold cross validations and an AUROC of 0.68 over the independent test, as shown in Figures 3 and 4. Recently, some methods have been developed for predicting the RNA–protein interface, including aPRBind [20], FastRNABindR [21], RNABindR v2 [22], BindN+ [33], and PPRInt [28]. The aPRBind [20] is a convolutional neural network-based method that uses sequence and structure information, while FastRNABindR [21], RNABindR v2 [22], BindN+ [33], and PPRInt [28] all adopt sequence-based features for interface prediction [20–22,28,33]. The performance of the independent test over the RB111 are listed in Table 2. The PRIP increased SN by 0.19 over

the aPRBind [20], by 0.06 over the FastRNABindR [21], by 0.04 over the RNABindR v2 [22], by 0.24 over the BindN+ [33], and by 0.19 over the PPRInt. On the other hand, the PRIP performed worst in terms of SP, ACC, and MCC. Apart from the PRIP, the best SN was 0.63, which was obtained by the RNABindR v2 [22]. This implied that it was challenging to correctly predict RNA–protein interfaces. Our method obtained a SN of 0.67. Table 3 lists the performance of the five-fold cross validations over the RB198. The same phenomenon was observed as in the independent test over the RB111. The PRIP obtained a better SN than the aPRBind [20].

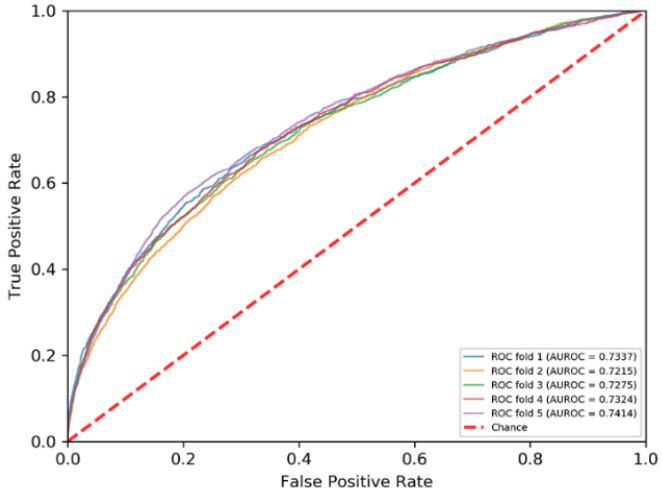

Figure 4. ROC curves of five-fold cross validation.

Table 2. PRIP was compared with existing methods on RB111.

Methods	SN	SP	ACC	MCC
PRIP	0.67	0.60	0.61	0.16
aPRBind [20]	0.48	0.90	0.86	0.32
FastRNABindR [21]	0.61	0.76	0.75	0.24
RNABindR v2 [22]	0.63	0.73	0.72	0.22
BindN+ [33]	0.43	0.87	0.84	0.24
PPRInt [28]	0.48	0.79	0.76	0.18

Table 3. The average results of the five-fold cross-validation studies on RB198 when compared to aPRBind.

Methods	SN	SP	ACC	MCC
PRIP	0.73	0.60	0.62	0.23
aPRBind [20]	0.65	0.82	0.74	0.48

3.4. Analysis of Pattern of the RNA-Binding Interfaces

We used the word cloud generator to draw a word cloud diagram of positive samples. As shown in Figure 5, the characters R, L, K, and G are dominant in the positive samples. We further employed Two Sample Logo [73] to visualize the difference between RNA-binding and non-binding protein sequences. Two Sample Logo is a tool to calculate and visualize differences between two sets of aligned samples of amino acids. Due to its simplicity and effectiveness, Two Sample Logo has widely been applied to the analysis of sequence patterns, such as post-translational modification patterns [74–76]. As shown in Figure 6, the characters R, K, and G are enriched in the RNA-binding protein sequences, and the characters L, A, E, and V are depleted. The results are in agreement with the word cloud diagram (Figure 5). This might imply that RNA-binding interfaces were associated with the emergence of R, K and G.

Figure 5. Word cloud of positive samples.

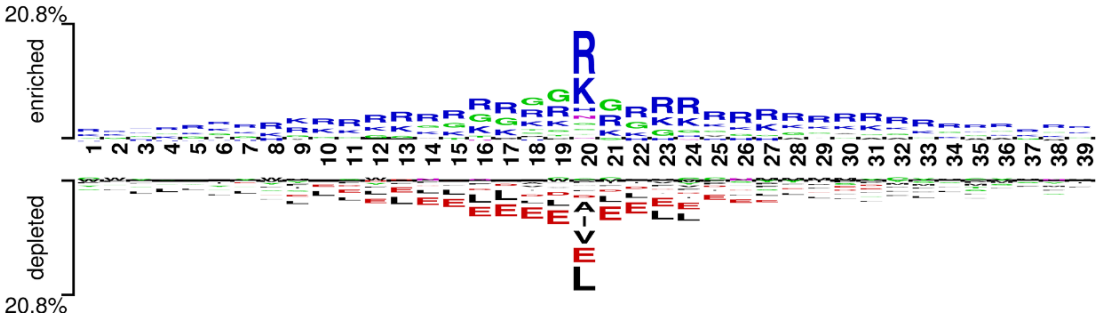

Figure 6. Two Sample logos: positive and negative samples correspond to the upper and lower parts respectively, and the height of residual letters is positively correlated with stacking order and frequency.

4. Discussion

We used only semantic embedding of protein sequences generated by the Word2vec to predict RNA-binding interfaces, and obtained competing performances with the state-of-the-art methods, including aPRBind [20], FastRNABindR [21], RNABindR v2 [22], BindN+ [33], and PPRInt [28]. This demonstrated that RNA-binding protein sequences were of semantics. The semantics of protein sequences have recently attracted attention from the molecular biology and bioinformatics communities [77]. For example, semantics were applied to detect remote evolutionary relationships [78,79], to predict protein sub-cellular localization [80], and to recognize protein–protein interactions [81]. Like natural language, biological sequences formed stable semantic relationships during the course of evolution. This is one of the reasons that our method obtained better performance in predicting RNA-binding interfaces.

In order to investigate the specificity of semantics, we generated four datasets of protein sequences by randomly altering 40%, 45%, 50%, and 55% of residues of RNA-binding protein sequences in the RB198. Each shuffled dataset was used as a corpus to pre-train Word2vec. The semantic relationship between words was defined as the cosine between the embedding of words, namely

$$\cos\theta = \frac{\sum_{i=1}^{n}(A_i \times B_i)}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}} \tag{21}$$

where $(A_1, A_2, \cdots, A_n)$ and $(B_1, B_2, \cdots, B_n)$ are the semantic embedding of the word A and B, respectively. Four shuffled datasets generated four stochastic semantic relationships for any two words. We used the RB198 as a corpus to pre-train the Word2vec, and we set the epochs to 100, 200, 300, and 400. We obtained four semantic embeddings of words. Using Equation (21), we computed true semantic relationships between any two words. We used student's test to investigate the difference in semantic relationships. As shown in Figure 7,

some parts of the semantic relationship are not of significant difference, while some are of significant difference (p-value < 0.05). For example, for L, the semantic relationship with three amino acids (T, M, and C) is of significant difference, while the semantic relationship of W with up to eight amino acids (E, A, V, G, F, Q, Y, C) is of significant difference. This indicated that some semantic relationships were specific to RNA-binding protein sequences.

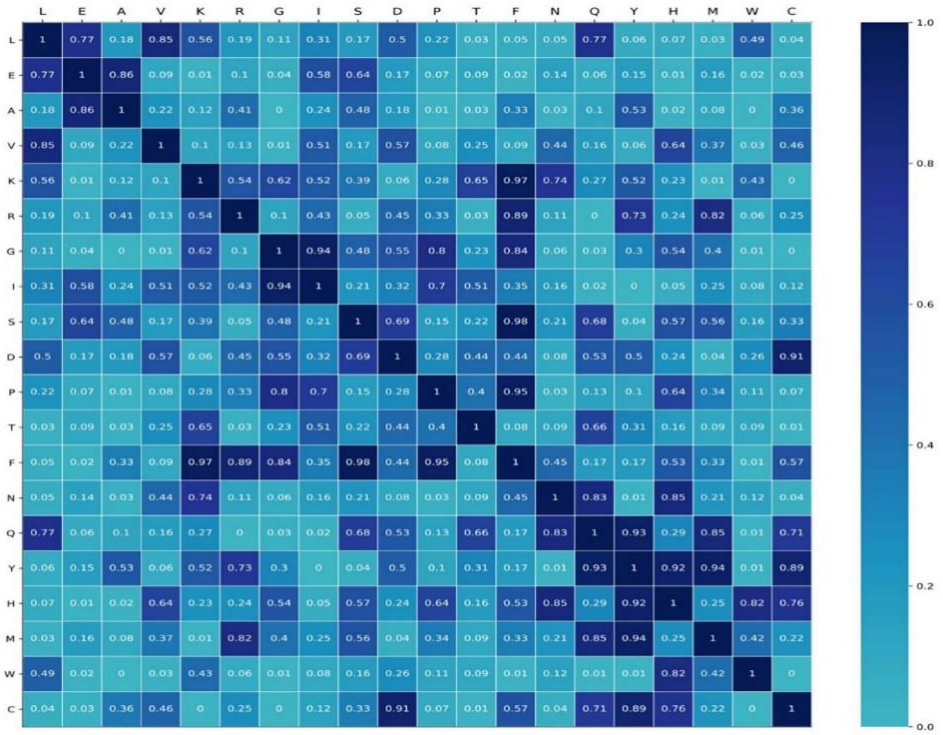

Figure 7. The p-value matrix by student's t-test.

As shown in Tables 2 and 3, the PRIP didn't show marked superiority over the state-of-the-art methods. This is because some negative segments are similar to the positive segments. For example, the protein sequence 'VERIFPL' has an RNA-binding interface, namely I. The protein sequence was assumed to be divided into three segments of five residues, 'VERIF', 'ERIFP', and 'RIFPL'. The interface was located at the center of the second segment, and thus it was a positive sample and the other two were negative ones. In fact, the two negative samples were very similar to the positive one, with only one different amino acid residue. Therefore, these semantics were too close to discriminate. After disrupting the negative fragments of the RNA-binding interface, we retrained the PRIP model (named it PRIP *) and repeated five-fold cross validation and the independent test. Table 4 shows the predictive performance. Obviously, SN and SP both increased, but the increase did not reach the expected value. There might be two reasons. One was that the original semantics of the negative samples were lost if we disrupted the sequence of all the negative samples. The other was that the motif of RNA-binding interfaces was quite complicated.

Table 4. Performances by the PRIP and the PRIP *.

	Methods	SN	SP	ACC	MCC
Five-fold cross validation	PRIP	0.73	0.60	0.62	0.23
	PRIP *	0.74	0.63	0.65	0.27
Independent test	PRIP	0.67	0.60	0.61	0.16
	PRIP *	0.69	0.62	0.63	0.18

As shown in Table 5, we conducted an analysis of two cases: 4V90_56 [82] and 3ULD_A [83]. The predictive performances are summarized in Table 5 for the RNABindRPlus [40], the PRIP, the PRIP *. Obviously, the PRIP * obtained the best SN, which is 0.16 more than that of the RNABindRPlus [40] and 0.12 more than that of the PRIP over the 4V90_56. Over the 3ULD_ A, the SN of the PRIP * is 0.13 more than that of the RNABindRPlus [40] and 0.20 more than that of the PRIP. Figure 8 illustrates the predicted structure of proteins in the gray cartoon.

Table 5. The predictive performances over the 4V90_56 and the 3ULD_A.

Protein	Methods	SN	SP	ACC	MCC
4V90_56	RNABindRPlus [40]	0.79	0.82	0.81	0.61
	PRIP	0.83	0.57	0.69	0.41
	PRIP *	0.95	0.65	0.78	0.62
3ULD_ A	RNABindRPlus [40]	0.80	0.87	0.86	0.53
	PRIP	0.73	0.67	0.68	0.26
	PRIP *	0.93	0.71	0.74	0.42

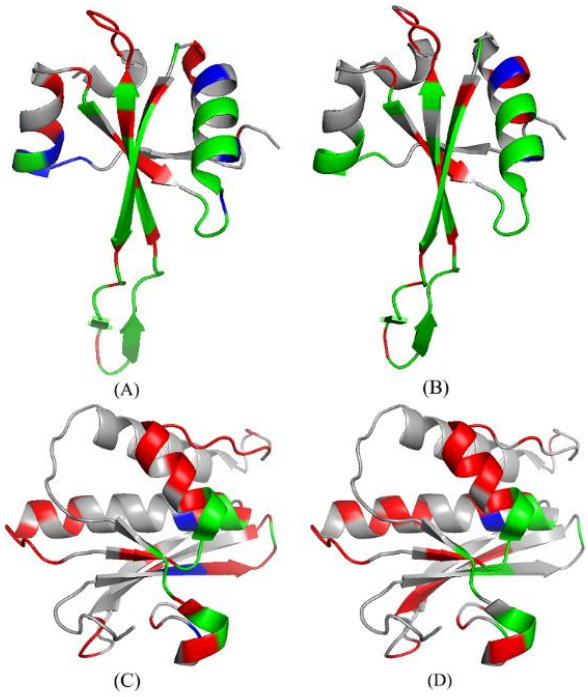

Figure 8. The protein structures of the predicted results of 4V90_56 and 3ULD_ A in PRIP and PRIP *. (**A**,**B**) represent the prediction results of 4V90_56 on PRIP and PRIP *, respectively. (**C**,**D**) represent the prediction results of 3ULD_ A on PRIP and PRIP *, respectively. Green, red, and blue represent TP, FP, and FN, respectively.

5. Conclusions

RNA–protein interactions play key roles in the regulation of many cellular processes and are increasingly becoming a hot topic. Although many computational methods have been presented in the past decades, it is still a challenging task to precisely and cheaply detect RNA-binding interfaces. We presented a sequence semantics-based method to predict RNA-binding interfaces. Compared with the state-of-the-art methods, the presented method learned the hidden relations between words in the context. This method is helpful to explore the mechanism of RNA–protein interactions from a semantics point of view.

Author Contributions: G.H. conceived the concept and methodology and wrote the manuscript. Y.W. and Y.L. (Yuewu Liu) conceived the methodology; Y.L. (You Li) collected the dataset, implemented

the methodology and software, and wrote the draft. J.L. wrote the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Scientific Research Fund of Hunan Provincial Education Department (21A0466, 19A215), by the Natural Science Foundation of Hunan Province (2020JJ4034), by the open project of Hunan Key Laboratory for Computation and Simulation in Science and Engineering (2019LCESE03), and by the Shaoyang University Innovation Foundation for Postgraduate (CX2021SY031).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data and source code can be found here: https://github.com/Good-Ly/PRIP.git (Accessed on: 4 February 2022).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Fabian, M.R.; Sonenberg, N.; Filipowicz, W. Regulation of mRNA translation and stability by microRNAs. *Annu. Rev. Biochem.* **2010**, *79*, 351–379. [CrossRef] [PubMed]
2. Hogan, D.J.; Riordan, D.P.; Gerber, A.P.; Herschlag, D.; Brown, P.O. Diverse RNA-binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system. *PLoS Biol.* **2008**, *6*, e255. [CrossRef] [PubMed]
3. Licatalosi, D.D.; Darnell, R.B. RNA processing and its regulation: Global insights into biological networks. *Nat. Rev. Genet.* **2010**, *11*, 75–87. [CrossRef] [PubMed]
4. Lorković, Z.J. Role of plant RNA-binding proteins in development, stress response and genome organization. *Trends Plant Sci.* **2009**, *14*, 229–236. [CrossRef]
5. Lukong, K.E.; Chang, K.-W.; Khandjian, E.W.; Richard, S. RNA-binding proteins in human genetic disease. *Trends Genet.* **2008**, *24*, 416–425. [CrossRef]
6. Verduci, L.; Tarcitano, E.; Strano, S.; Yarden, Y.; Blandino, G. CircRNAs: Role in human diseases and potential use as biomarkers. *Cell Death Dis.* **2021**, *12*, 1–12. [CrossRef]
7. Gebauer, F.; Schwarzl, T.; Valcárcel, J.; Hentze, M.W. RNA-binding proteins in human genetic disease. *Nat. Rev. Genet.* **2021**, *22*, 185–198. [CrossRef]
8. Saunus, J.M.; French, J.D.; Edwards, S.L.; Beveridge, D.J.; Hatchell, E.C.; Wagner, S.A.; Stein, S.R.; Davidson, A.; Simpson, K.J.; Francis, G.D. Posttranscriptional regulation of the breast cancer susceptibility gene BRCA1 by the RNA binding protein HuR. *Cancer Res.* **2008**, *68*, 9469–9478. [CrossRef]
9. Esteller, M. Non-coding RNAs in human disease. *Nat. Rev. Genet.* **2011**, *12*, 861–874. [CrossRef]
10. Khalil, A.M.; Rinn, J.L. RNA–protein interactions in human health and disease. *Semin. Cell Dev. Biol.* **2011**, *22*, 359–365. [CrossRef]
11. Van Roosbroeck, K.; Pollet, J.; Calin, G.A. miRNAs and long noncoding RNAs as biomarkers in human diseases. *Expert Rev. Mol. Diagn.* **2013**, *13*, 183–204. [CrossRef] [PubMed]
12. Guo, Q.; Dammer, E.B.; Zhou, M.; Kundinger, S.R.; Gearing, M.; Lah, J.J.; Levey, A.I.; Shulman, J.M.; Seyfried, N.T. Targeted Quantification of Detergent-Insoluble RNA-Binding Proteins in Human Brain Reveals Stage and Disease Specific Co-aggregation in Alzheimer's Disease. *Front. Mol. Neurosci.* **2021**, *14*, 623659. [CrossRef] [PubMed]
13. Tan, L.; Yu, J.-T.; Hu, N.; Tan, L. Non-coding RNAs in Alzheimer's disease. *Mol. Neurobiol.* **2013**, *47*, 382–393. [CrossRef] [PubMed]
14. Schonrock, N.; Götz, J. Decoding the non-coding RNAs in Alzheimer's disease. *Cell. Mol. Life Sci.* **2012**, *69*, 3543–3559. [CrossRef]
15. Schultz, C.W.; Preet, R.; Dhir, T.; Dixon, D.A.; Brody, J.R. Understanding and targeting the disease-related RNA binding protein human antigen R (HuR). *Wiley Interdiscip. Rev. RNA* **2020**, *11*, e1581. [CrossRef]
16. Shi, X.; Sun, M.; Liu, H.; Yao, Y.; Kong, R.; Chen, F.; Song, Y. A critical role for the long non-coding RNA GAS5 in proliferation and apoptosis in non-small-cell lung cancer. *Mol. Carcinog.* **2015**, *54*, E1–E12. [CrossRef]
17. Congrains, A.; Kamide, K.; Oguro, R.; Yasuda, O.; Miyata, K.; Yamamoto, E.; Kawai, T.; Kusunoki, H.; Yamamoto, H.; Takeya, Y. Genetic variants at the 9p21 locus contribute to atherosclerosis through modulation of ANRIL and CDKN2A/B. *Atherosclerosis* **2012**, *220*, 449–455. [CrossRef]
18. Ke, A.; Doudna, J.A. Crystallization of RNA and RNA–protein complexes. *Methods* **2004**, *34*, 408–414. [CrossRef]
19. Scott, L.G.; Hennig, M. RNA structure determination by NMR. *Bioinformatics* **2008**, *452*, 29–61.
20. Liu, Y.; Gong, W.; Zhao, Y.; Deng, X.; Zhang, S.; Li, C. aPRBind: Protein-RNA interface prediction by combining sequence and I-TASSER model-based structural features learned with convolutional neural networks. *Bioinformatics* **2021**, *37*, 937–942. [CrossRef]
21. El-Manzalawy, Y.; Abbas, M.; Malluhi, Q.; Honavar, V. FastRNABindR: Fast and accurate prediction of protein-RNA Interface residues. *PLoS ONE* **2016**, *11*, e0158445.

22. Walia, R.R.; Caragea, C.; Lewis, B.A.; Towfic, F.; Terribilini, M.; El-Manzalawy, Y.; Dobbs, D.; Honavar, V. Protein-RNA interface residue prediction using machine learning: An assessment of the state of the art. *BMC Bioinform.* **2012**, *13*, 1–20. [CrossRef] [PubMed]
23. Liu, Z.-P.; Wu, L.-Y.; Wang, Y.; Zhang, X.-S.; Chen, L. Prediction of protein-RNA binding sites by a random forest method with combined features. *Bioinformatics* **2010**, *26*, 1616–1622. [CrossRef] [PubMed]
24. Carson, M.B.; Langlois, R.; Lu, H. NAPS: A residue-level nucleic acid-binding prediction server. *Nucleic Acids Res.* **2010**, *38*, W431–W435. [CrossRef] [PubMed]
25. Cheng, C.-W.; Su, E.C.-Y.; Hwang, J.-K.; Sung, T.-Y.; Hsu, W.-L. Predicting RNA-binding sites of proteins using support vector machines and evolutionary information. *BMC Bioinform.* **2008**, *9*, 1–19. [CrossRef]
26. Jeong, E.; Chung, I.F.; Miyano, S. A neural network method for identification of RNA-interacting residues in protein. *Genome Inform.* **2004**, *15*, 105–116.
27. Jeong, E.; Miyano, S. A weighted profile based method for protein-RNA interacting residue prediction. In *Transactions on Computational Systems Biology IV*; Springer: Berlin/Heidelberg, Germany, 2006; pp. 123–139.
28. Kumar, M.; Gromiha, M.M.; Raghava, G.P.S. Prediction of RNA binding sites in a protein using SVM and PSSM profile. *Proteins Struct. Funct. Bioinform.* **2008**, *71*, 189–194. [CrossRef]
29. Ma, X.; Guo, J.; Wu, J.; Liu, H.; Yu, J.; Xie, J.; Sun, X. Prediction of RNA-binding residues in proteins from primary sequence using an enriched random forest model with a novel hybrid feature. *Proteins Struct. Funct. Bioinform.* **2011**, *79*, 1230–1239. [CrossRef]
30. Spriggs, R.V.; Murakami, Y.; Nakamura, H.; Jones, S. Protein function annotation from sequence: Prediction of residues interacting with RNA. *Bioinformatics* **2009**, *25*, 1492–1497. [CrossRef]
31. Terribilini, M.; Lee, J.H.; Yan, C.; Jernigan, R.L.; Honavar, V.; Dobbs, D. Prediction of RNA binding sites in proteins from amino acid sequence. *RNA* **2006**, *12*, 1450–1462. [CrossRef]
32. Wang, C.-C.; Fang, Y.; Xiao, J.; Li, M. Identification of RNA-binding sites in proteins by integrating various sequence information. *Amino Acids* **2011**, *40*, 239–248. [CrossRef]
33. Wang, L.; Huang, C.; Yang, M.Q.; Yang, J.Y. BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features. *BMC Syst. Biol.* **2010**, *4*, 1–9. [CrossRef] [PubMed]
34. Wang, L.; Brown, S.J. Prediction of RNA-binding residues in protein sequences using support vector machines. In Proceedings of the 2006 International Conference of the IEEE Engineering in Medicine and Biology Society, New York, NY, USA, 30 August–3 September 2006; pp. 5830–5833.
35. Kim, O.T.; Yura, K.; Go, N. Amino acid residue doublet propensity in the protein-RNA interface and its application to RNA interface prediction. *Nucleic Acids Res.* **2006**, *34*, 6450–6460. [CrossRef] [PubMed]
36. Maetschke, S.R.; Yuan, Z. Exploiting structural and topological information to improve prediction of RNA-protein binding sites. *BMC Bioinform.* **2009**, *10*, 341. [CrossRef]
37. Pérez-Cano, L.; Fernández-Recio, J. Optimal protein-RNA area, OPRA: A propensity-based method to identify RNA-binding sites on proteins. *Proteins Struct. Funct. Bioinform.* **2010**, *78*, 25–35. [CrossRef] [PubMed]
38. Towfic, F.; Caragea, C.; Gemperline, D.C.; Dobbs, D.; Honavar, V. Struct-NB: Predicting protein-RNA binding sites using structural features. *Int. J. Data Min. Bioinform.* **2010**, *4*, 21–43. [CrossRef] [PubMed]
39. Zhao, H.; Yang, Y.; Zhou, Y. Structure-based prediction of RNA-binding domains and RNA-binding sites and application to structural genomics targets. *Nucleic Acids Res.* **2011**, *39*, 3017–3025. [CrossRef]
40. Walia, R.R.; Xue, L.C.; Wilkins, K.; El-Manzalawy, Y.; Dobbs, D.; Honavar, V. RNABindRPlus: A predictor that combines machine learning and sequence homology-based methods to improve the reliability of predicted RNA-binding residues in proteins. *PLoS ONE* **2014**, *9*, e97725.
41. Pan, X.; Yang, Y.; Xia, C.Q.; Mirza, A.H.; Shen, H.B. Recent methodology progress of deep learning for RNA–protein interaction prediction. *Wiley Interdiscip. Rev. RNA* **2019**, *10*, e1544. [CrossRef]
42. Zhang, W.; Qu, Q.; Zhang, Y.; Wang, W. The linear neighborhood propagation method for predicting long non-coding RNA–protein interactions. *Neurocomputing* **2018**, *273*, 526–534. [CrossRef]
43. Zhang, H.; Ming, Z.; Fan, C.; Zhao, Q.; Liu, H. A path-based computational model for long non-coding RNA-protein interaction prediction. *Genomics* **2020**, *112*, 1754–1760. [CrossRef] [PubMed]
44. Adjeroh, D.; Allaga, M.; Tan, J.; Lin, J.; Jiang, Y.; Abbasi, A.; Zhou, X. Feature-Based and String-Based Models for Predicting RNA-Protein Interaction. *Molecules* **2018**, *23*, 697. [CrossRef] [PubMed]
45. Liu, R.; Hu, J. HemeBIND: A novel method for heme binding residue prediction by combining structural and sequence information. *BMC Bioinform.* **2011**, *12*, 1–14. [CrossRef] [PubMed]
46. Altschul, S.F.; Madden, T.L.; Schaffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389–3402. [CrossRef] [PubMed]
47. Zhang, W.; Yue, X.; Tang, G.; Wu, W.; Huang, F.; Zhang, X. SFPEL-LPI: Sequence-based feature projection ensemble learning for predicting LncRNA-protein interactions. *PLoS Comput. Biol.* **2018**, *14*, e1006616. [CrossRef]
48. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. *arXiv* **2013**, arXiv:1301.3781.

49. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 5 December–8 December 2013; pp. 3111–3119.
50. Alshemali, B.; Kalita, J. Improving the reliability of deep neural networks in NLP: A review. *Knowl. Based Syst.* **2020**, *191*, 105210. [CrossRef]
51. Tsukiyama, S.; Hasan, M.M.; Fujii, S.; Kurata, H. LSTM-PHV: Prediction of human-virus protein-protein interactions by LSTM with word2vec. *Brief. Bioinform.* **2021**, *22*, bbab228. [CrossRef]
52. Hamid, M.-N.; Friedberg, I. Identifying antimicrobial peptides using word embedding with deep recurrent neural networks. *Bioinformatics* **2019**, *35*, 2009–2016. [CrossRef]
53. Wu, C.; Gao, R.; Zhang, Y.; De Marinis, Y. PTPD: Predicting therapeutic peptides by deep learning and word2vec. *BMC Bioinform.* **2019**, *20*, 456. [CrossRef]
54. Sharma, R.; Shrivastava, S.; Kumar Singh, S.; Kumar, A.; Saxena, S.; Kumar Singh, R. Deep-ABPpred: Identifying antibacterial peptides in protein sequences using bidirectional LSTM with word2vec. *Brief. Bioinform.* **2021**. [CrossRef] [PubMed]
55. Ofer, D.; Brandes, N.; Linial, M. The language of proteins: NLP, machine learning & protein sequences. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 1750–1758. [PubMed]
56. Iuchi, H.; Matsutani, T.; Yamada, K.; Iwano, N.; Sumi, S.; Hosoda, S.; Zhao, S.; Fukunaga, T.; Hamada, M. Representation learning applications in biological sequence analysis. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 3198–3208. [CrossRef] [PubMed]
57. Song, B.; Li, Z.; Lin, X.; Wang, J.; Wang, T.; Fu, X. Pretraining model for biological sequence data. *Brief. Funct. Genom.* **2021**, *20*, 181–195. [CrossRef] [PubMed]
58. Lewis, B.A.; Walia, R.R.; Terribilini, M.; Ferguson, J.; Zheng, C.; Honavar, V.; Dobbs, D. PRIDB: A protein-RNA interface database. *Nucleic Acids Res.* **2010**, *39*, D277–D282. [CrossRef] [PubMed]
59. Puton, T.; Kozlowski, L.; Tuszynska, I.; Rother, K.; Bujnicki, J.M. Computational methods for prediction of protein-RNA interactions. *J. Struct. Biol.* **2012**, *179*, 261–268. [CrossRef]
60. Goldberg, Y.; Levy, O. word2vec Explained: Deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv* **2014**, arXiv:1402.3722.
61. Rong, X. word2vec parameter learning explained. *arXiv* **2014**, arXiv:1411.2738.
62. Rehurek, R.; Sojka, P. Software framework for topic modelling with large corpora. In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, Valleta, Malta, 19 May–21 May 2010; pp. 45–50.
63. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016; pp. 785–794.
64. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
65. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
66. Menard, S. *Applied Logistic Regression Analysis*; Sage: Hongkong, China, 2002; Volume 106.
67. Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
68. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. Lightgbm: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Processing Syst.* **2017**, *30*, 3146–3154.
69. Bhasin, M.; Raghava, G.P. Classification of nuclear receptors based on amino acid composition and dipeptide composition. *J. Biol. Chem.* **2004**, *279*, 23262–23266. [CrossRef] [PubMed]
70. Saravanan, V.; Gautham, N. Harnessing computational biology for exact linear B-cell epitope prediction: A novel amino acid composition-based feature descriptor. *Omics A J. Integr. Biol.* **2015**, *19*, 648–658. [CrossRef] [PubMed]
71. Chen, Z.; Zhao, P.; Li, F.; Leier, A.; Marquez-Lago, T.T.; Wang, Y.; Webb, G.I.; Smith, A.I.; Daly, R.J.; Chou, K.-C. iFeature: A python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics* **2018**, *34*, 2499–2502. [CrossRef]
72. Cao, R.; Wang, M.; Bin, Y.; Zheng, C. DLFF-ACP: Prediction of ACPs based on deep learning and multi-view features fusion. *PeerJ* **2021**, *9*, e11906. [CrossRef]
73. Vacic, V.; Iakoucheva, L.M.; Radivojac, P. Two Sample Logo: A graphical representation of the differences between two sets of sequence alignments. *Bioinformatics* **2006**, *22*, 1536–1537. [CrossRef]
74. Huang, G.; Zheng, Y.; Wu, Y.-Q.; Han, G.-S.; Yu, Z.-G. An information entropy-based approach for computationally identifying histone lysine butyrylation. *Front. Genet.* **2020**, *10*, 1325. [CrossRef]
75. Xiang, Q.; Feng, K.; Liao, B.; Liu, Y.; Huang, G. Prediction of Lysine Malonylation Sites Based on Pseudo Amino Acid. *Comb. Chem. High Throughput Screen.* **2017**, *20*, 622–628. [CrossRef]
76. Xu, Y.; Wang, Z.; Li, C.; Chou, K.-C. iPreny-PseAAC: Identify C-terminal cysteine prenylation sites in proteins by incorporating two tiers of sequence couplings into PseAAC. *Med. Chem.* **2017**, *13*, 544–551. [CrossRef]
77. Lipton, R.J.; Marr, T.G.; Welsh, J.D. Computational approaches to discovering semantics in molecular biology. *Proc. IEEE* **1989**, *77*, 1056–1060. [CrossRef]
78. Dong, Q.W.; Wang, X.L.; Lin, L. Application of latent semantic analysis to protein remote homology detection. *Bioinformatics* **2006**, *22*, 285–290. [CrossRef] [PubMed]
79. Melvin, I.; Weston, J.; Noble, W.S.; Leslie, C. Detecting remote evolutionary relationships among proteins by large-scale semantic embedding. *PLoS Comput. Biol.* **2011**, *7*, e1001047. [CrossRef] [PubMed]

80. Chang, J.M.; Su, E.C.Y.; Lo, A.; Chiu, H.S.; Sung, T.Y.; Hsu, W.L. PSLDoc: Protein subcellular localization prediction based on gapped-dipeptides and probabilistic latent semantic analysis. *Proteins: Struct. Funct. Bioinform.* **2008**, *72*, 693–710. [CrossRef]
81. Wang, Y.; You, Z.-H.; Yang, S.; Li, X.; Jiang, T.-H.; Zhou, X. A high efficient biological language model for predicting protein–protein interactions. *Cells* **2019**, *8*, 122. [CrossRef]
82. Chen, Y.; Feng, S.; Kumar, V.; Ero, R.; Gao, Y.-G. Structure of EF-G–ribosome complex in a pretranslocation state. *Nat. Struct. Mol. Biol.* **2013**, *20*, 1077–1084. [CrossRef]
83. Gan, J.H.; Abdur, R.; Huang, Z. RNA/DNA Hybrid in Complex with RNase H catalytic Domain Mutant D132N. 2011. Available online: https://www.rcsb.org/structure/3ULD (accessed on 5 January 2022).