Computational method for predicting Self-Interactions Protein using Recurrent Neural Network from Protein Evolutionary Information

Ji-Yong An (ajy@cumt.edu.cn)

Research article

Keywords: SIPs, Recurrent Neural Network, Scale Invariant Feature Transform, PSSM

Posted Date: February 17th, 2020

DOI: https://doi.org/10.21203/rs.2.23742/v1

Computational method for predicting Self-Interactions Protein using Recurrent Neural Network from Protein Evolutionary Information

Ji-Yong An[1,*]
(ajy@cumt.edu.cn)
[1]School of Computer Science and Technology, China University of Mining and Technology
Xuzhou Jiangsu 21116, China
Corresponding author mail: ajy@cumt.edu.cn

Abstract: Self-interactions Protein (SIPs) play crucial roles in biological activities of organisms. Many high-throughput methods can be used to identify SIPs. However, these methods are both time-consuming and expensive. How to develop effective computational approaches for identifying SIPs is a challenging task. In the paper, we presented a novelty computational method called RRN-SIFT, which combines the Recurrent Neural Network (RNN) with Scale Invariant Feature Transform (SIFT) to predict SIPs based on protein evolutionary information. The main advantage of the proposed RNN-SIFT model is that it used SIFT for extracting key feature by exploring the evolutionary information embedded in PSI-BLAST-constructed position specific scoring matrix (PSSM) and employed RNN classifier to carry out classification based on extracted features. Extensive experiments show that the RRN-SIFT obtained average accuracy of 94.34% and 97.12% on *yeast* and *human* dataset. We also compared our performance with the Back Propagation Neural Network (BPNN), the state-of-the-art support vector machine (SVM) and other exiting methods. By comparing with experimental results, the performance of RNN-SIFT is significantly better than those of the BPNN, SVM and other previous methods in the domain. Therefore, we can come to the conclusion that the proposed RNN-SIFT model is useful tools and can execute incredibly well for predicting SIPs, as well as other bioinformatics tasks. In order to facilitate widely studies and encourage future proteomics research, a freely available web server called RNN-SIFT-SIPs was developed, and is available at http://219.219.62.123:8888/RNNSIFT/ and includes source code and SIPs datasets.

Key words: SIPs, Recurrent Neural Network, Scale Invariant Feature Transform, PSSM

1. Introduction

Protein-protein interactions (PPIs) prediction revealed multiple roles in many important biological activity. However, an interesting research problem regarding whether proteins can interact with their partner. Self-interactions protein (SIPs) is being considered as a special type of PPIs, which refers to more than two copies of the protein can interact with each other and are the same copies of the protein and can be represented by the same gene. This might bring about the formation of homo-oligomer problem. Many recent studies have shown that SIPs play a vital role in various cellular physiological functions and the evolution process of protein-protein interaction networks (PPINs) [1-3]. Therefore, whether a protein can self-interact for interpretation its functions is very important. The research on SIPs can provide a certain help for a far better understanding the regulation of protein function and the molecular mechanisms involved in biological activity and the underlying cellular and genetic disease mechanisms. Many studies have been conducted for the homo-oligomerization that is a vital function for biological activity and plays an absolutely essential role in a wide range of biological processes, such as, signal transduction, gene expression regulation, enzyme activation and immune response[4-8]. In addition, it has been

demonstrated by many previous studies that the diversity function of proteins can be variously extended without increasing the length of genome through SIPs. SIPs can also provide some help in improving the protein stability and preventing the protein denaturation by reducing its surface area [9, 10]. Therefore, it is becoming more and more important to develop reliable and highly effective computational approaches based protein sequence for predicting SIPs.

As always, a large number of researches have been devoted to develop reliable and highly effective computational approaches to predict PPIs. Gao et al [11] proposed a novelty computational method called RF-AC, which combined the Rotation Forest (RF) classifier with Auto covariance (AC) approach based PSSM. Huang et al [12] presented a new computational approach, which used weighted sparse representation (WSRC) as classifier and employed global encoding (GE) as feature extraction method for predicting PPIs. Pan et al [13]proposed a novelty latent dirichlet allocation-random forest model (LDA-RF) for predicting human PPIs based on protein primary sequences, which is strong ability for processing large-scale datasets by using LDA-RF model. Zhang et al [14] proposed a novel approach based on protein sequence that used Random Tree and Genetic Algorithm for predicting PPIs, which obtained good prediction results. Yang et al [15] presented a new approach that used Local descriptors to represent protein sequence and employed the k-nearest neighbors for carrying out classification. Guo et al [16] adopted autocorrelation feature extraction technique for generating feature vectors and used the SVM classifier to identify PPIs. An et al [17] proposed a classification algorithm of compound kernel function RVM based on gray wolf optimization algorithm and K-fold cross Validation, which fully consider the special features of local and global of protein-protein interactions position and obtained good predicting results. An et al [18] proposed a feature extraction approach based on local protein sequence PSSM matrix coding and serial multi-feature Fusion. The method can capture protein-protein interaction information of continuous and discontinuous for protein sequence by using the local protein sequence PSSM matrix coding; much key feature information contained protein sequences can be integrated through employing serial multi-feature Fusion. These methods usually explored the correlational information between protein pairs, such as, coevolution, co-localization and co-expression. However, this information is not enough for predicting SIPs. In addition, the PPIs datasets do not contain the PPIs between the same protein partners. With all these reasons, it is not fit for predicting SIPs by using these computational approaches. In the previous study, Liu et al [1] proposed a method integrating multiple representative known properties to create a prediction mode called as SLIPPER to predict SIPs. As far as we know, a number of recent studies have been reported about PPIs, which may also be related to SIPs [19-21]. However, there is an obviously drawback that cannot deal with the proteins not covering the current human interatomic by using these methods. Due to all the reasons hereinbefore, it is an urgent work at present for developing efficient computational approaches for predicting SIPs.

In the study, we proposed a novelty computational method named RRN-SIFT, which combines the Recurrent Neural Network (RNN) with Scale Invariant Feature Transform (SIFT) to predict SIPs based on protein evolutionary information. The major advantage of the proposed RNN-SIFT model is that it used SIFT for extracting key feature by exploring the evolutionary information embedded in PSI-BLAST-constructed position specific scoring matrix (PSSM) and employed RNN classifier to carry out classification based on extracted features. Extensive experiments show that the RRN-SIFT obtained average accuracy of 94.34% and 97.12% on yeast and human dataset. We also compared our performance with the Back Propagation Neural Network (BPNN), the state-of-

the-art support vector machine (SVM) and other exiting methods. By comparing with experimental results, the performance of RNN-SIFT is significantly better than those of the BPNN, SVM and other previous methods in the domain. Therefore, we can come to the conclusion that the proposed RNN-SIFT model is useful tools and can execute incredibly well for predicting SIPs, as well as other bioinformatics tasks.

2. Materials and Methodology

2.1. Dataset

The UniProt database contains 20,199 curated *human* protein sequences [22]. The PPIs datasets can be downloaded from different databases, including DIP [23], BioGRID[24], IntAct[25], InnateDB [26] and MatrixDB [27]. The PPIs data was constructed in the paper that only contains the same two interaction protein sequence, whose interaction type was defined as 'direct interaction' in relevant databases. As a result, 2994 human Protein Self-interactions protein sequences were obtained. In order to verify the performance of the RNN-SIFT model, we constructed the experimental datasets by using as following three steps[28]: (1) the protein sequences whose length less than 50 residues and longer than 5000 residues were removed from the whole human proteome;(2) we selected the SIPs data to create the positive dataset, which must satisfy one of the following conditions: (a) it has been detected for the Self-interactions by at least two kinds of large scale experiments or one small-scale experiment; (b) the protein has been defined as homooligomer (including homodimer and homodimers) in UniProt; (c) it has been reported by at least two publications for the Self-interactions;(3) for constructing the negative dataset, we removed all types of SIPs from the whole human proteome (including proteins annotated as 'direct interaction' and more extensive 'physical association') and UniProt database. Consequently, we selected 15,938 non-SIPs as negatives samples and 1441 SIPs as positives samples for creating the human dataset[28]. In addition, we also used the same strategy to construct the *yeast* dataset that contains 5511 negative and 710 positive samples[28].

2.2. Feature Extraction Method

2.2.1 Position Specific Scoring Matrix (PSSM)

Position Specific Scoring Matrix (PSSM) contains not only the position information but also the evolution information of protein sequence. As a result, the PSSM is used to extract the evolutionary information in the paper. Using Position Specific Iterated BLAST (PSI-BLAST) [29] to convert each sequence into a PSSM. Assuming the length of a given protein sequence is *L*, its PSSM can be expressed as an $L \times 20$ matrix. Figure 1 shows the schematic of a PSSM.

$$PSSM = \begin{bmatrix} P_{1,1} & P_{1,2} & P_{1,3} & \cdots & P_{1,20} \\ P_{2,1} & P_{2,2} & P_{2,3} & \cdots & P_{2,20} \\ \vdots & P_{i,j} & \vdots & \vdots & \vdots \\ P_{L,1} & P_{L,2} & P_{L,3} & \cdots & P_{L,20} \end{bmatrix}$$

Figure 1 the schematic of a PSSM

Where *L* represents the length of a given sequence, 20 are the number of 20 amino acids, and P_{ij} represents the score of the j_{th} amino acid in the i_{th} position for the query sequence. The P_{ij} can be greater than 0, less than 0 or equal to 0. If P_{ij} is greater than 0, it means that the i_{th} amino acid is easily mutated into the j_{th} amino acid during the evolution process, and a larger value indicates a higher mutation probability. Conversely, if P_{ij} is less than 0, the position is conservative

and the probability of mutation is small. Smaller P_{ij} are more conservative. To extract evolutionary information from protein sequences, each SIP's sequence was converted into a PSSM by using the PSI-BLAST tool. To obtain highly and widely homologous sequences, PSI_BLAST's e-value parameter was set to 0.001 and three iterations were selected.

2.2.2 Scale Invariant Feature Transform (SIFT)

Scale Invariant Feature Transform (SIFT) is an image descriptor developed by David Lowe, which was used to match and recognition image-based[30, 31]. The original SIFT descriptor was calculated from the image intensities around interesting locations in the image domain which can be named interest points or key points. These interest points are obtained from scale-space extrema of differences-of-Gaussians (DOG) within a difference-of Gaussians pyramid. Lindeberg[32, 33] proposed a new method for finding out interest points by using the SIFT approach. This method can be viewed as a variation of a scale-adaptive blob detection approach, where blobs with associated scale levels are detected from scale-space extrema of the scale-normalized Laplacian. The scale-normalized Laplacian is normalized with respect to the scale level in scale-space and is defined as：

$$\begin{aligned}\nabla^2_{norm}L(x,y^{\cdot},s) &= s\big(L_{xx}+L_{yy}\big) = s\left(\frac{\partial^2 L}{\partial x^2}+\frac{\partial^2 L}{\partial y^2}\right) \\ &= s\nabla^2(G((x,y^{\cdot},s)*f(x,y))\end{aligned} \quad (1)$$

For obtaining the maximum value of the DOG image under different scale magnifications, the smoothed image values of a given original image is convolved with Gaussian kernels of different widths by using SIFT algorithm, a scale-variable Gaussian function is defined as follow:

$$G((x,y^{\cdot},s) = \frac{1}{2\pi s}e^{-(x^2+y^2)}/(2s) \quad (2)$$

These Gaussian-blurred images are grouped according to their scale magnification, so the number of Gaussian blur images processed in each group is the same. At this time, the DOG image can be obtained by subtracting two adjacent Gaussian blurred images in the same group. The difference-of-Gaussians operator constitutes an approximation of the Laplacian operator of different widths, where denotes the standard deviation and the variance of the Gaussian kernel. The difference-of-Gaussians operator constitutes an approximation of the Laplacian operator is defined as follow:

$$DOG((x,y^{\cdot},s) = L(x,y^{\cdot},s+\Delta s) - L(x,y^{\cdot},s)) \approx \frac{\Delta s}{2}\nabla^2 L(x,y^{\cdot},s) \quad (3)$$

Which by the implicit normalization of the differences-of-Gaussian responses, as obtained by a self-similar distribution of scale levels $\sigma_{i+1} = k\sigma_i$ used by Lowe, also constitutes an approximation of the scale-normalized Laplacian with $\Delta s\nabla^2 L = (k^2-1)t\nabla^2 L = (k^2-1)\nabla^2_{norm}L$ thus implying

$$DOG((x,y^{\cdot},s) \approx \frac{(k^2-1)}{2}\nabla^2_{norm}L(x,y^{\cdot},s) \quad (4)$$

After the DOG image is obtained, the maximum and minimum values can be found and is referred to as key points in the DOG images. In order to quickly find the key points, each pixel of the DOG image will be compared with the eight pixels around itself and nine pixels at the same position in the same group of the DOG images at adjacent scales. The maximum and minimum values of these pixels are called key points. As a result, the critical point detection of SIFT algorithm is actually a variant of Blob detection, which use Laplacian to compute the maximum values in each magnification space. The Gaussian difference can be approximated as the result of Laplace operator

operation. SIFT employs the concept of "scale space" to capture features at multiple scale levels or image resolutions, which not only increases the number of available features but also makes the method highly tolerant to scale changes.

In the paper, we assumed that each PSSM is an image matrix. As a result, we used SIFT feature extraction method to generate feature vectors and its dimensional is 128. The technology roadmap of the proposed method is shown in Figure 2.

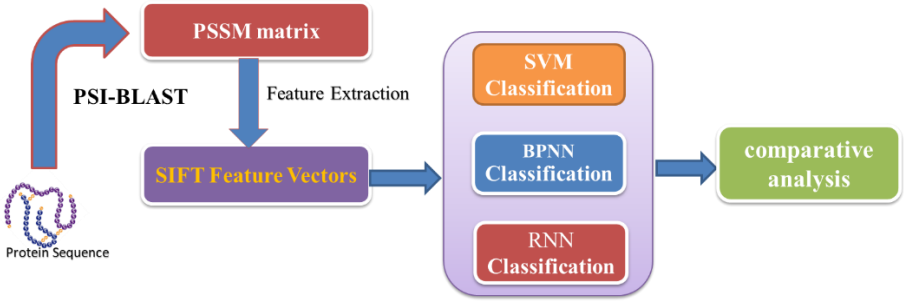

Figure 2 the technology roadmap of the proposed method

2.3 Recurrent Neural Network (RNN)

Recurrent Neural Network (RNN) is used to solve the problem that the input training samples is a continuous sequence and the length of the sequence is different, such as the problem based on time series. The basic neural network only establishes weight connections between layers. The biggest difference of RNN is that the weight connections also established between layers of neurons [34-36]. The structure of RNN is as follows:

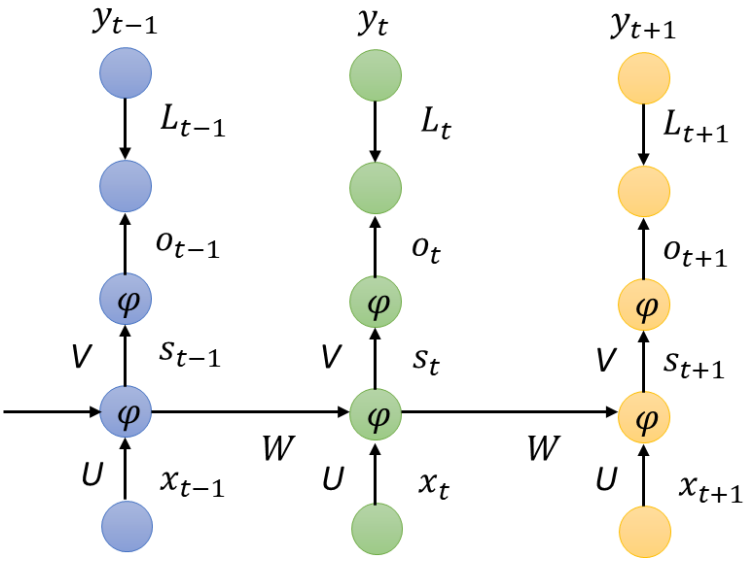

Figure 3 the structure of RNN

It can be seen from Figure 3 that the output of RNN at any moment is related to the current input and the previous output. RNN's forward propagation is a combination of multiplication, addition and set operations. It is well known that *t* moment of a given ordered sequence will lead to computation of the hidden layer *t* times. The current state of hidden layer $h(t)$ is determined by the current input $x(t)$ and the output $h(t-1)$ of the previous layer. The mathematical description is as follows:

$$s(t) = Ux(t) + Wh(t-1) + b \quad (5)$$

$$h(t) = \sigma\big(s(t)\big) = \sigma(Ux(t) + Wh(t-1) + b) \quad (6)$$

Where σ represents activation function. The output of the current hidden layer can be calculated by using the following function:

$$\sigma(t) = Vh(t) + c \quad (7)$$

The Softmax function can be used to carry out classification and output the final prediction probability value, which is shown as follow:

$$y_p = \sigma\big(o(t)\big) = \sigma(Vh(t) + c) \quad (8)$$

Here, the loss function of y_p is different from y. In practice, we can select different loss functions according to the need of the different problem, such as, the log loss function, the square loss function, and so on. The loss function of the RNN model at moment t can be expressed as follows:

$$Loss_t = -[y_t\, ln(o_t) + (y_t - 1)\, ln(1 - o_t)] \quad (9)$$

The loss function (global loss) of the RNN model at all moments N can be expressed as follows:

$$Loss = \sum_{t=1}^{N} Loss_t = -\sum_{t=1}^{N}[y_t\, ln(o_t) + (y_t - 1)\, ln(1 - o_t)] \quad (10)$$

The gradient of three parameters U, V, and W of the global loss can be defined as follows:

$$\begin{aligned} \frac{\partial L}{\partial V} &= \sum_{t=1}^{N} (o_t - y_t) \times s_t \\ \frac{\partial L}{\partial U} &= \sum_{t=1}^{N} \sum_{k=1}^{t} \frac{\partial L_t}{\partial s_k^*} \times x_k^T \\ \frac{\partial L}{\partial U} &= \sum_{t=1}^{N} \sum_{k=1}^{t} \frac{\partial L_t}{\partial s_k^*} \times s_{k-1}^T \end{aligned} \quad (11)$$

The most commonly used method for optimization problems is the gradient descent. In the paper, the gradient update for the three parameters can be expressed as follows:

$$\begin{aligned} V: &= V - \eta \times \partial L/\partial V \\ U: &= U - \eta \times \partial L/\partial U \\ W: &= W - \eta \times \partial L/\partial W \end{aligned} \quad (12)$$

The major advantage of RNN model in learning nonlinear sequential data is well-known and has been utilized in language modeling and sequential labeling. In consideration of SIPs dataset is also a kind of nonlinear sequence data, so we used RNN model to predict SIPs in the study. The prediction flowchart of RNN-SIFT model is displayed in Figure 4.

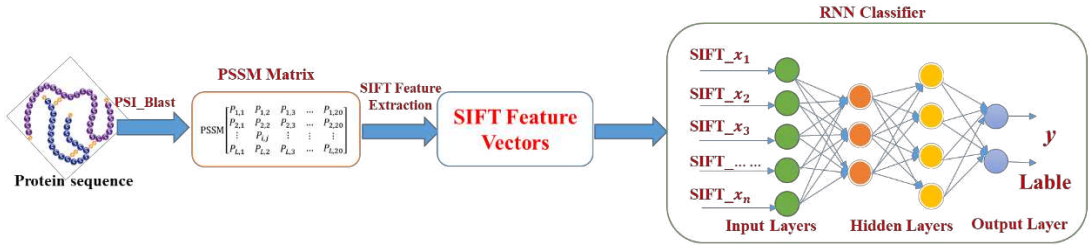

Figure 4 the prediction flowchart of RNN-SIFT

2.4. Performance Evaluation

In the paper, we employed the following measures to assess the performance of RNN-SIFT.

$$\mathrm{Ac} = \frac{TP + TN}{TP + FP + TN + FN} \quad (13)$$

$$\mathrm{Sn} = \frac{TP}{TP + TN} \quad (14)$$

$$\mathrm{Sp} = \frac{TN}{FP + TN} \quad (15)$$

$$\mathrm{Pe} = \frac{TP}{FP + TP} \quad (16)$$

$$\mathrm{Mcc} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}} \quad (17)$$

Where Ac is Accuracy, Sn represents Sensitivity, Sp is specificity， Pe represents Precision and Mcc is Matthews's correlation coefficient. TP and TN represent the number of true interacting and true non-interacting pairs that were correctly predicted, respectively. FP and FN is the count of true non-interacting pairs and true interacting pairs falsely predicted, respectively. In addition, we used Receiver Operating Curve (ROC) to further evaluate the performance of RNN-SIFT in the experiment.

3. Results and Discussion

3.1. Performance of the proposed RNN-SIFT model

In the experiment, we used *yeast a*nd *human* datasets to evaluate the proposed RNN-SIFT model. Generally overfitting will affect experimental results. Therefore, we divided the whole datasets into the training datasets and independent test datasets for preventing overfitting. Specifically, we split the *yeast* dataset into 6 parts, and selected 5 parts of them as the training set and the remaining dataset selected as independent test dataset. The *human* dataset was also processed by using the same strategy. Meanwhile, fivefold cross-validation tests was employed to evaluate the performance of the RNN-SIFT for fair comparison and several parameters of the RNN model were optimized through using the grid search for ensuring fairness. Here, we set up learning rate=0.001, training step = 1000 and hidden units = 200. Table 1-2 shows the experimental results of the proposed RNN-SIFT model on *yeast* and *human* dataset.

As can be seen from table 1, the proposed RNN-SIFT model obtained good experimental results on *yeast* dataset. The result of average accuracy 94.34 %, average Sensitivity 67.12%, Precision 79.79% and MCC 71.61% were achieved in the experiments on fivefold cross-validation tests. Similarly, another promising finding from Table 2 was that the RNN-SIFT also achieved better prediction results on *human* dataset, whose average accuracy, sensitivity, precision, and MCC are 97.12%, 83.70%, 85.24% and 79.35% respectively. As a result, the proposed RNN-SIFT model has high value in research.

The good experimental results for predicting SIPs are mainly attributed to use the SIFT feature extraction method and RNN classifier. The main advantage of the RNN-SIFT model is that SIFT method can extract key evaluation feature from PSSM and RNN classifier has the advantage of processing sequence data. As discussed, this is mainly due to the following three reasons: (1) PSSM contains not only the position information but also the evolution information of protein sequence, and retains plenty of prior information. This make it possible to contains a number of key features can be extracted. (2) SIFT uses the concept of "scale space" to capture features at multiple scale levels, which not only increases the number of available features but also makes the method highly tolerant to scale changes. This makes it possible for extracting the evolutionary information embedded in PSSM and capturing self-protein interaction information. (3) Recurrent neural networks have some characteristics in memory, parameter sharing and Turing completeness, so which provide an advantage for learning based on the nonlinear characteristics of sequences. Therefore, RNN is use to carry out classification for predicting SIPs. The results demonstrate two

things. First, SIFT method is very suitable for extracting self-protein sequence feature. Second, the RNN classifier performs well for predicting SIPs, giving good results.

Table 1 Fivefold cross validation results shown using RNN-SIFT model on *yeast*

Testing set	Ac (%)	Sn (%)	Pe (%)	Mcc (%)
1	94.11	64.29	70.79	66.31
2	94.59	70.83	80.19	73.79
3	93.15	62.40	76.47	67.51
4	94.98	65.29	88.76	74.45
5	94.88	72.80	82.73	75.97
Average	**94.34±0.74**	**67.12±4.46**	**79.79±6.73**	**71.61±4.38**

Table 2 Fivefold cross validation results shown using RNN-SIFT model on *human*

Testing set	Ac (%)	Sn (%)	Pe (%)	Mcc (%)
1	97.10	73.27	84.57	77.76
2	97.24	74.89	83.94	80.79
3	96.89	74.59	82.73	79.25
4	96.75	72.45	82.53	77.46
5	97.63	76.37	87.02	81.51
Average	**97.12±0.34**	**83.70±1.15**	**85.24±2.92**	**79.35±1.79**

3.2. Comparison with the Method of BPNN-based and SVM-based

It is interesting to note that the RNN-SIFT model is very suitable for predicting SIPs and can obtain good prediction results. However, to further evaluate the performance of RNN-SIFT model, we compared the results of the RNN classifier with those of the Back Propagation Neural Network (BPNN) classifier and the Support Vector Machine (SVM) classifier by using the same SIFT approach on *yeast* and *human* datasets, respectively. In order to ensure fair comparison, several parameter settings of BPNN were optimized by employing grid search approach. Specifically, the epochs, the eta, the BS and the WS of BPNN are set to 100, 0.006, 0.5 and 0.7. Similarly, by using the same strategy as described above, the RBF kernel parameters of the SVM were optimized, where c is 0.5 and g is 10.8 and other parameters takes the default value. In addition, SVM classifier used the LIBSVM tool [37] to carry out classification.

Table 3-6 below shows the experimental results of BPNN-SIF and SVM-SIFT on *yeast* and *human* dataset, respectively. Meanwhile the comparison of ROC Curves on yeast and human dataset between RNN, BPNN and SVM are shown in Figure 5-6 below respectively. As outlined in Table 3-4, the BPNN-SIFT model achieved 91.31% average accuracy and the SVM-SIFT model obtained 89.58% average accuracy on yeast dataset. Similarly, as can be seen from table 5-6, the results of average accuracy 93.84% and 91.79% are obtained by the BPNN-SIFT model and the SVM-SIFT model on human dataset, respectively. When comparing our results to those of BPNN-SIFT and SVM-SIFT, it must be pointed out that the performance of RNN classifier is significantly better than that of the other two classifiers. At the same time, from Figure 5 and Figure 6, the ROC curves of RNN classifier are also significantly better than that of the other two classifiers. A major reason for good prediction results is that Self-protein sequence is nonlinear sequence data and RNN classifier have some characteristics in memory, parameter sharing and Turing completeness and can provide

an advantage for learning based on the nonlinear characteristics of sequences. From the above analysis, the paper comes to the conclusion that the proposed RNN-SIFT model is useful tools for identifying SIPs, as well as other bioinformatics tasks.

Table 3 Fivefold cross validation results shown by using BPNN-SIFT model on *yeast*

Testing set	Ac (%)	Sn (%)	Pe (%)	Mcc (%)
1	92.86	60.80	75.25	66.11
2	91.70	49.59	70.59	58.06
3	90.64	47.20	65.56	54.81
4	89.86	48.33	57.43	52.31
5	91.47	47.12	61.11	57.84
Average	**91.31±1.13**	**50.61±5.78**	**65.99±7.15**	**57.82±5.20**

Table 4 Fivefold cross validation results shown by using SVM-SIFT model on *yeast*

Testing set	Ac (%)	Sn (%)	Pe (%)	Mcc (%)
1	89.57	31.63	81.58	49.68
2	90.05	35.33	85.19	55.48
3	89.08	30.40	79.63	48.96
4	90.02	33.88	87.23	52.62
5	89.21	30.12	71.45	46.58
Average	**89.58±0.45**	**33.27±2.26**	**81.02±6.12**	**50.66±3.45**

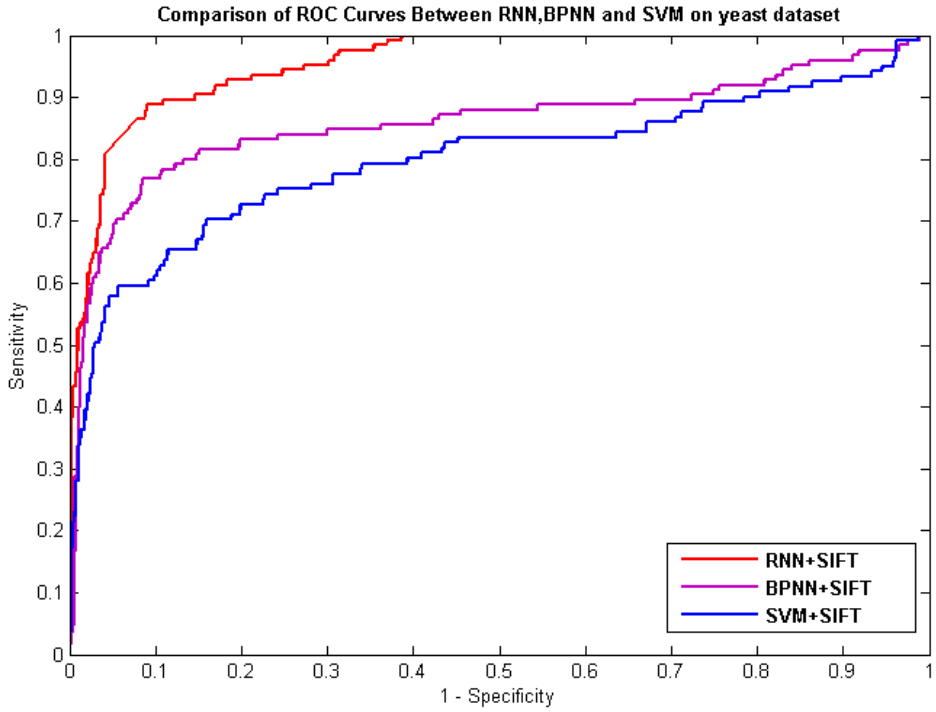

Figure 5 Comparison of ROC curves between RNN, BPNN and SVM on *yeast* dataset.

Table 5 Fivefold cross validation results shown by using BPNN-SIFT model on *human*

Testing set	Ac (%)	Sn (%)	Pe (%)	Mcc (%)
1	94.10	51.61	83.33	68.17
2	94.41	58.30	85.80	73.43
3	93.44	50.41	81.79	66.65
4	92.51	45.28	79.55	62.22
5	94.75	54.85	89.04	68.65
Average	**93.84±0.89**	**52.09±4.89**	**83.90±3.67**	**67.82±4.03**

Table 6 Fivefold cross validation results shown by using SVM-SIFT model on *human*

Testing set	Ac (%)	Sn (%)	Pe (%)	Mcc (%)
1	92.57	38.21	83.87	57.68
2	91.80	33.33	88.89	52.62
3	90.73	28.00	85.37	47.27
4	91.70	33.88	87.23	51.72
5	92.18	36.00	87.83	56.98
Average	**91.79±0.69**	**33.88±3.81**	**86.64±2.01**	**53.23±4.22**

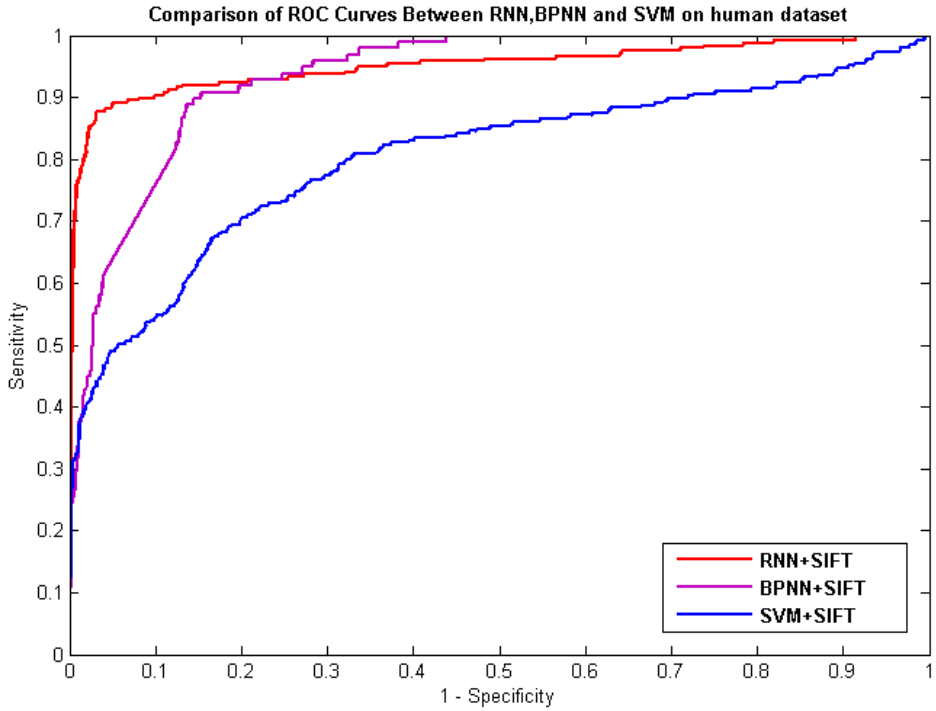

Figure 6 Comparison of ROC curves between RNN, BPNN and SVM on *human* dataset.

3.3. Comparison with Other Methods

To go a step further and validate the performance of the proposed RNN-SIFT model, we compare the prediction results of RNN-SIFT model with those of the previous methods ,such as ,SLIPPER[38],CRS[28], SPAR[28] , DXECPPI , PPIevo [39] and LocFuse [40]. Table7-8 shown a detailed comparison results on yeast and human dataset. It can be seen from Table 7 that the average accuracy of RNN-SIFT is obviously higher than those of the other six approaches on yeast

dataset. Similarity, Table 8 displays the prediction accuracy obtained RNN-SIFT model is also significantly better than those of the other six methods on human dataset. A similar conclusion was reached by comparing the results from Table 7-8 that the proposed RNN-SIFT model has an excellent prediction capability and can be used for quality predicting SIPs. This is a result of using a robust RNN classifier and an effectively SIFT feature extraction technique. These comparison results is further evidence that the RNN-SIFT is suit for predicting SIPs.

Table 7 Comparison results between RNN-SIFT and other methods on *yeast* dataset

Model	Ac (%)	Sp (%)	Sn (%)	MCC
SLIPPER[38]	71.90	72.18	69.72	0.2842
PPIevo[39]	66.28	87.46	60.14	0.1801
LocFuse[40]	66.66	68.10	55.49	0.1577
CRS[28]	72.69	74.37	59.58	0.2368
SPAR[28]	76.96	80.02	53.24	0.2484
Proposed method	**94.34**	**79.79**	**67.12**	**0.7161**

Table 8 Comparison results between RNN-SIFT and other methods on *human* dataset

Model	Ac (%)	Sp (%)	Sn (%)	MCC
SLIPPER[38]	91.10	95.06	47.26	0.4197
PPIevo[39]	78.04	25.82	87.83	0.2082
LocFuse[40]	80.66	80.50	50.83	0.2026
CRS[28]	91.54	96.72	34.17	0.3633
SPAR[28]	92.09	97.40	33.33	0.3836
Proposed method	**97.12**	**85.24**	**83.70**	**0.7935**

4. Conclusion

In the study, we proposed a novelty computational method named RRN-SIFT, which combines the Recurrent Neural Network (RNN) with Scale Invariant Feature Transform (SIFT) to predict SIPs based on protein evolutionary information. Extensive experiments show that the RRN-SIFT obtained average accuracy of 94.34% and 97.12% on yeast and human dataset. We also compared our performance with the Back Propagation Neural Network (BPNN), the state-of-the-art support vector machine (SVM) and other exiting methods. By comparing with experimental results, the performance of RNN-SIFT is significantly better than those of the BPNN, SVM and other previous methods in the domain. This is mainly due to the following three reasons: (1) PSSM contains not only the position information but also the evolution information of protein sequence, and retains plenty of prior information. This make it possible to contains a number of key features can be extracted. (2) SIFT uses the concept of “scale space” to capture features at multiple scale levels, which not only increases the number of available features but also makes the method highly tolerant to scale changes. This makes it possible for extracting the evolutionary information embedded in PSSM and capturing self-protein interaction information. (3) Self-protein sequence is nonlinear sequence data and RNN have some characteristics in memory, parameter sharing and Turing completeness and can provide an advantage for learning based on the nonlinear characteristics of

sequences. Therefore, we can come to the conclusion that the proposed RNN-SIFT model is useful tools and can execute incredibly well for predicting SIPs, as well as other bioinformatics tasks.

Declarations

Availability of data and material: In this study, our experimental datasets contain *yeast* and *human* dataset, which can be obtained from the publicly available DIP [23], BioGRID[24], IntAct[25], InnateDB [26] and MatrixDB [27].

Competing interests: The authors declare no conflict of interest.

Funding: This work is supported by 'the Fundamental Research Funds for the Central Universities (2019XKQYMS88)'.

Author Contributions: Ji-Yong An conceived the algorithm, carried out analyses, prepared the data sets, carried out experiments, wrote the manuscript, and approved the final manuscript.

Acknowledgments: The authors would like to thank all the guest editors and anonymous reviewers for their constructive advices.

References

1. **Self Protein**. *Encyclopedia of Genetics Genomics Proteomics & Informatics* 2008.
2. Brun VL, Friess W, Schultz-Fademrecht T, Muehlau S, Garidel P: **Lysozyme-lysozyme self-interactions as assessed by the osmotic second virial coefficient: Impact for physical protein stabilization**. *Biotechnology Journal* 2010, **4**(9):1305-1319.
3. Zhai JX, Cao T-J, An J-Y, Bian Y-T: **Highly accurate prediction of protein self-interactions by incorporating the average block and PSSM information into the general PseAAC**. *Journal of Theoretical Biology*:S0022519317303752.
4. Baisamy L, Jurisch N, Diviani D: **Leucine zipper-mediated homo-oligomerization regulates the Rho-GEF activity of AKAP-Lbc**. *Journal of Biological Chemistry* 2005, **280**(15):15405-15412.
5. Hattori T, Ohoka N, Inoue Y, Hayashi H, Onozaki K: **C/EBP family transcription factors are degraded by the proteasome but stabilized by forming dimer**. *Oncogene* 2003, **22**(9):1273-1280.
6. Katsamba P, Carroll K, Ahlsen G, Bahna F, Vendome J, Posy S, Rajebhosale M, Price S, Jessell TM, Ben-Shaul A: **Linking molecular affinity and cellular specificity in cadherin-mediated adhesion**. *Proceedings of the National Academy of Sciences of the United States of America* 2009, **106**(28):11594-11599.
7. Koike R, Kidera A, Ota M: **Alteration of oligomeric state and domain architecture is essential for functional transformation between transferase and hydrolase with the same scaffold**. *Protein Science A Publication of the Protein Society* 2009, **18**(10):2060.
8. Woodcock JM, Murphy J, Stomski FC, Berndt MC, Lopez AF: **The dimeric versus monomeric status of 14-3-3zeta is controlled by phosphorylation of Ser58 at the dimer interface**. *Journal of Biological Chemistry* 2003, **278**(38):36323.
9. Marianayagam NJ, Sunde M, Matthews JM: **The power of two: protein dimerization in biology**. *Trends in Biochemical Sciences* 2004, **29**(11):618-625.
10. An JY, Zhang L, Zhou Y, Zhao Y-J, Wang D-FJJoC: **Computational methods using weighed-extreme learning machine to predict protein self-interactions with protein evolutionary information**. *Journal of Cheminformatics*, **9**(1):47.
11. Gao ZG, Lei W, Shi-Xiong X, Zhu-Hong Y, Xin Y, International ZYJBR: **Ens-PPI: A Novel Ensemble**

Classifier for Predicting the Interactions of Proteins Using Autocovariance Transformation from PSSM. *Biomed Research International*, **2016**:1-8.

12. Huang YA, You Z-H, Chen X, Chan K, Luo XJBB: **Sequence-based prediction of protein-protein interactions using weighted sparse representation model combined with global encoding**. *Bmc Bioinformatics*, **17**(1):184.
13. Pan XY, Zhang YN, Shen HBJJoPR: **Large-Scale Prediction of Human Protein-Protein Interactions from Amino Acid Sequence Based on Latent Topic Features**. *Journal of Proteome Research* 2010, **9**(10):4992-5001.
14. Lei Z: **Sequence-Based Prediction of Protein-Protein Interactions Using Random Tree and Genetic Algorithm**. In: *International Conference on Intelligent Computing: 2012*.
15. Yang L, Xia JF, Gui J: **Prediction of protein-protein interactions from protein sequence using local descriptors**. *Protein & Peptide Letters* 2010, **17**(9):1085.
16. Guo Y, Yu L, Wen Z, Li M: **Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences**. *Nucleic Acids Research* 2008, **36**(9):3025.
17. An JY, Zhou Y, Zhao YJ, Yan ZJ: **An Efficient Feature Extraction Technique Based on Local Coding PSSM and Multifeatures Fusion for Predicting Protein-Protein Interactions**. *Evol Bioinform* 2019, **15**:10.
18. An JY, You ZH, Zhou Y, Wang DF: **Sequence-based Prediction of Protein-Protein Interactions Using Gray Wolf Optimizer-Based Relevance Vector Machine**. *Evol Bioinform* 2019, **15**:10.
19. Jia J, Liu Z, Xiao X, Liu B, Chou KC: **iPPI-Esml: An ensemble classifier for identifying the interactions of proteins by incorporating their physicochemical properties and wavelet transforms into PseAAC**. *Journal of Theoretical Biology* 2015, **377**:47-56.
20. Jia J, Liu Z, Xiao X, Liu B, Chou KC: **Identification of protein-protein binding sites by incorporating the physicochemical properties and stationary wavelet transforms into pseudo amino acid composition**. *Journal of Biomolecular Structure & Dynamics* 2015:1-38.
21. Jia J, Liu Z, Xiao X, Liu B, Chou KC: **iPPBS-Opt: A Sequence-Based Ensemble Classifier for Identifying Protein-Protein Binding Sites by Optimizing Imbalanced Training Datasets**. *Molecules* 2015, **21**(1):E95.
22. Consortium UP: **UniProt: a hub for protein information**. *Nucleic Acids Research* 2014, **43**(D1):D204-212.
23. Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM, Eisenberg D: **DIP: the database of interacting proteins**. *Nucleic Acids Research* 2004, **32**(1):D449.
24. Livstone MS, Breitkreutz BJ, Stark C, Boucher L, Chatraryamontri A, Oughtred R, Nixon J, Reguly T, Rust J, Winter A: **The BioGRID Interaction Database**. 2011, **41**(Database issue):: D637–D640.
25. Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackescarter F, Campbell NH, Chavali G, Chen C, Deltoro N: **The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases**. *Nucleic Acids Research* 2014, **42**:358-363.
26. Breuer K, Foroushani AK, Laird MR, Chen C, Sribnaia A, Lo R, Winsor GL, Hancock REW, Brinkman FSL, Lynn DJ: **InnateDB: Systems biology of innate immunity and beyond - Recent updates and continuing curation**. *Nucleic Acids Research* 2013, **41**(Database issue):D1228.
27. Launay G, Salza R, Multedo D, Thierrymieg N, Ricardblum S: **MatrixDB, the extracellular matrix interaction database: updated content, a new navigator and expanded functionalities**. *Nucleic Acids Research* 2014, **43**(Database issue):321-327.

28. Liu X, Yang S, Li C, Zhang Z, Song J: **SPAR: a random forest-based predictor for self-interacting proteins with fine-grained domain information**. *Amino Acids* 2016, **48**(7):1655.
29. Gribskov M, Mclachlan AD, Eisenberg D: **Profile analysis: detection of distantly related proteins**. *Proceedings of the National Academy of Sciences of the United States of America* 1987, **84**(13):4355.
30. Lowe DG: **Object Recognition from Local Scale-Invariant Features**. In: *iccv: 1999*.
31. Lowe DG: **Distinctive Image Features from Scale-Invariant Keypoints**. *International Journal of Computer Vision* 2004, **60**(2):91---110.
32. Lindeberg, Tony: **SCALE-SPACE THEORY IN COMPUTER VISION**. 1994, **256**:349-382.
33. Lindeberg T: **Feature Detection with Automatic Scale Selection**. *International Journal of Computer Vision*, **30**(2):79-116.
34. Gregor K, Danihelka I, Graves A, Rezende DJ, Wierstra D: **DRAW: A Recurrent Neural Network For Image Generation**. *Computer Science* 2015:1462-1471.
35. Lukoševičius M, Jaeger H: **Reservoir computing approaches to recurrent neural network training**. 2009, **3**(3):127-149.
36. Donkers T, Loepp B, Ziegler J: **Sequential User-based Recurrent Neural Network Recommendations**. In: *RecSys '17: Proceedings of the the 11th ACM Conference on Recommender Systems: 2017*.
37. Chih-Chung, Chang, Chih-Jen, Lin: **LIBSVM: A library for support vector machines**.
38. Liu Z, Guo F, Zhang J, Wang J, Lu L, Li D, He F: **Proteome-wide prediction of self-interacting proteins based on multiple properties**. *Molecular & Cellular Proteomics Mcp* 2013, **12**(6):1689.
39. Zahiri J, Yaghoubi O, Mohammad-Noori M, Ebrahimpour R, Masoudi-Nejad A: **PPIevo: Protein-Protein Interaction Prediction from PSSM Based Evolutionary Information**. *Genomics* 2013, **102**(4):237-242.
40. Zahiri J, Mohammad-Noori M, Ebrahimpour R, Saadat S, Bozorgmehr JH, Goldberg T, Masoudi-Nejad A: **LocFuse: Human protein–protein interaction prediction via classifier fusion using protein localization information**. *Qrevchemsoc* 2014, **104**(6):496-503.

Figures

$$\text{PSSM}\begin{bmatrix} P_{1,1} & P_{1,2} & P_{1,3} & \cdots & P_{1,20} \\ P_{2,1} & P_{2,2} & P_{2,3} & \cdots & P_{2,20} \\ \vdots & P_{i,j} & \vdots & \vdots & \vdots \\ P_{L,1} & P_{L,2} & P_{L,3} & \cdots & P_{L,20} \end{bmatrix}$$

Figure 1

the schematic of a PSSM

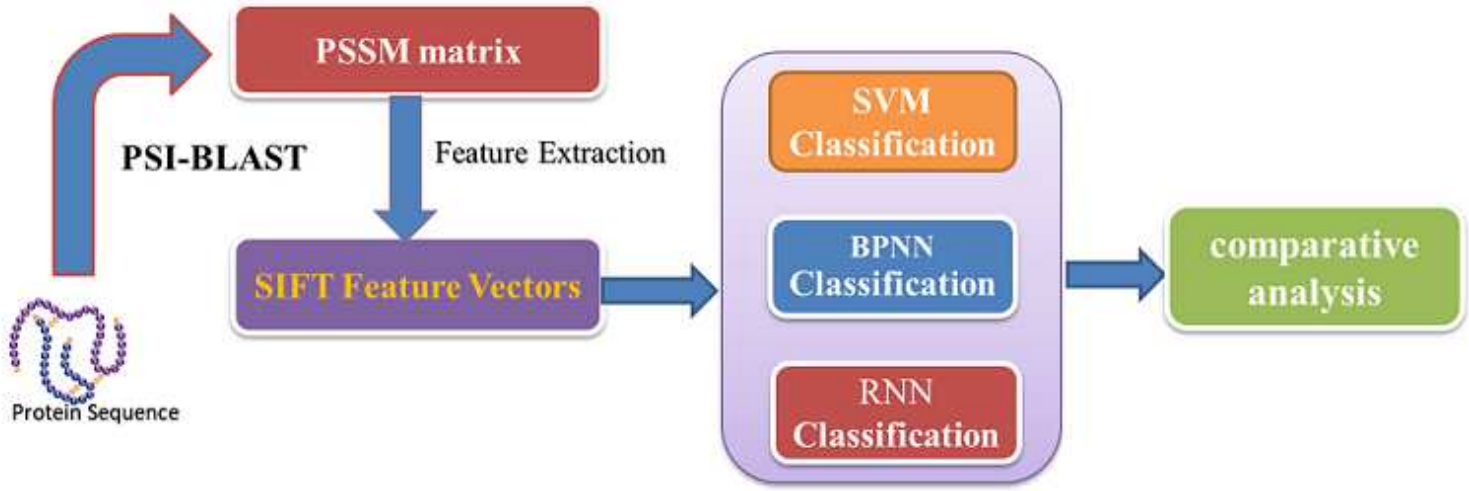

Figure 2

the technology roadmap of the proposed method

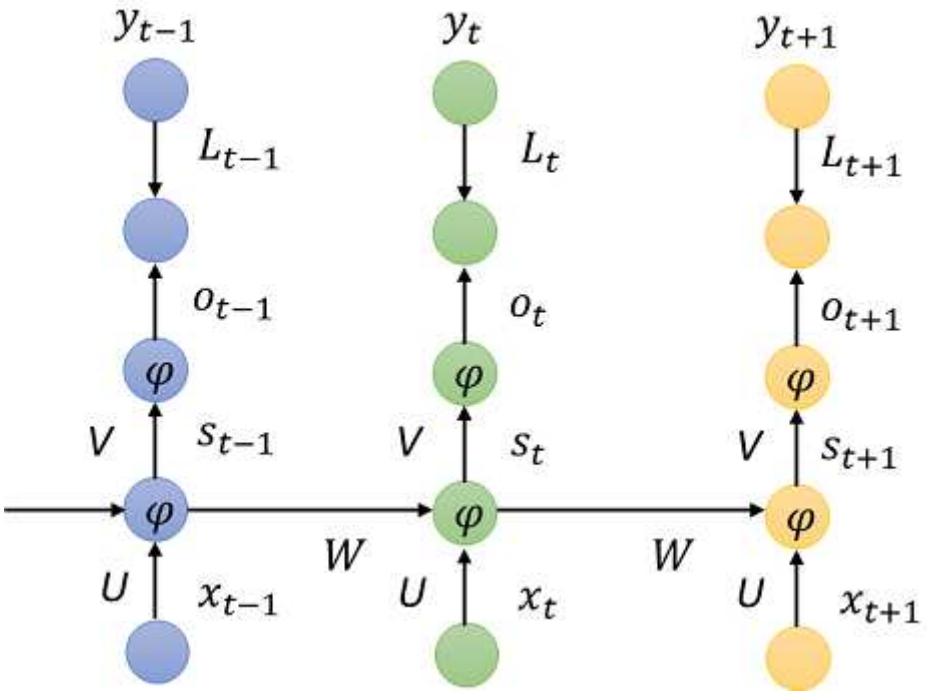

Figure 3

the structure of RNN

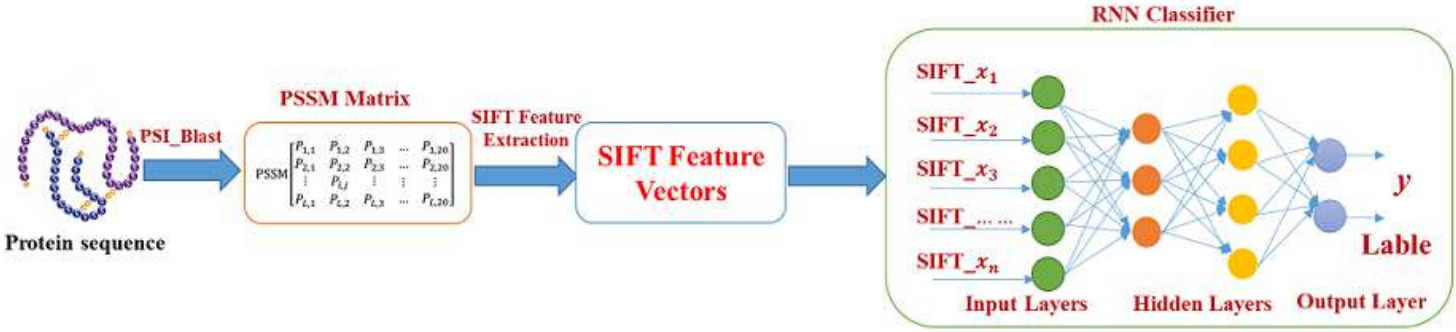

Figure 4

the prediction flowchart of RNN-SIFT

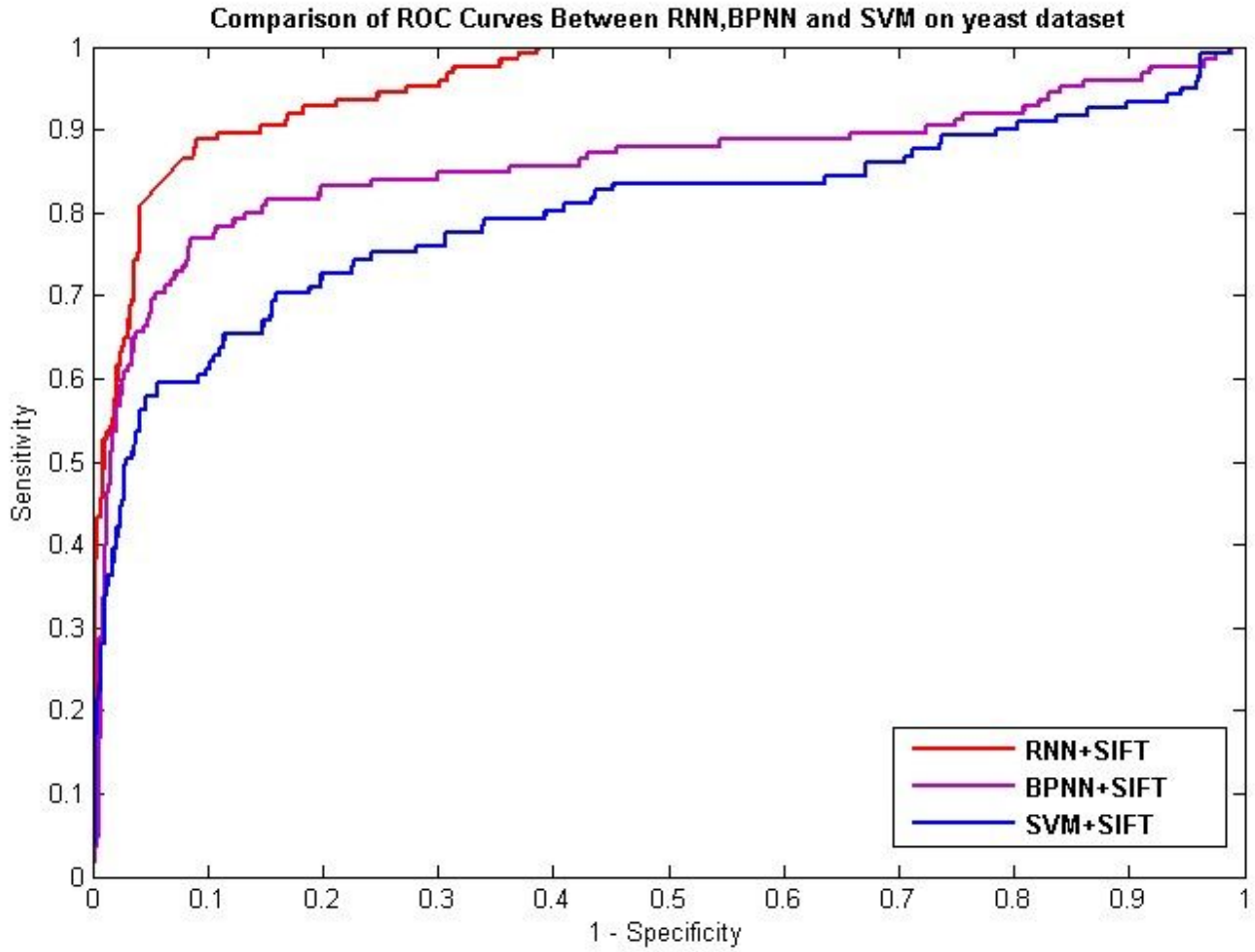

Figure 5

Comparison of ROC curves between RNN, BPNN and SVM on yeast dataset.

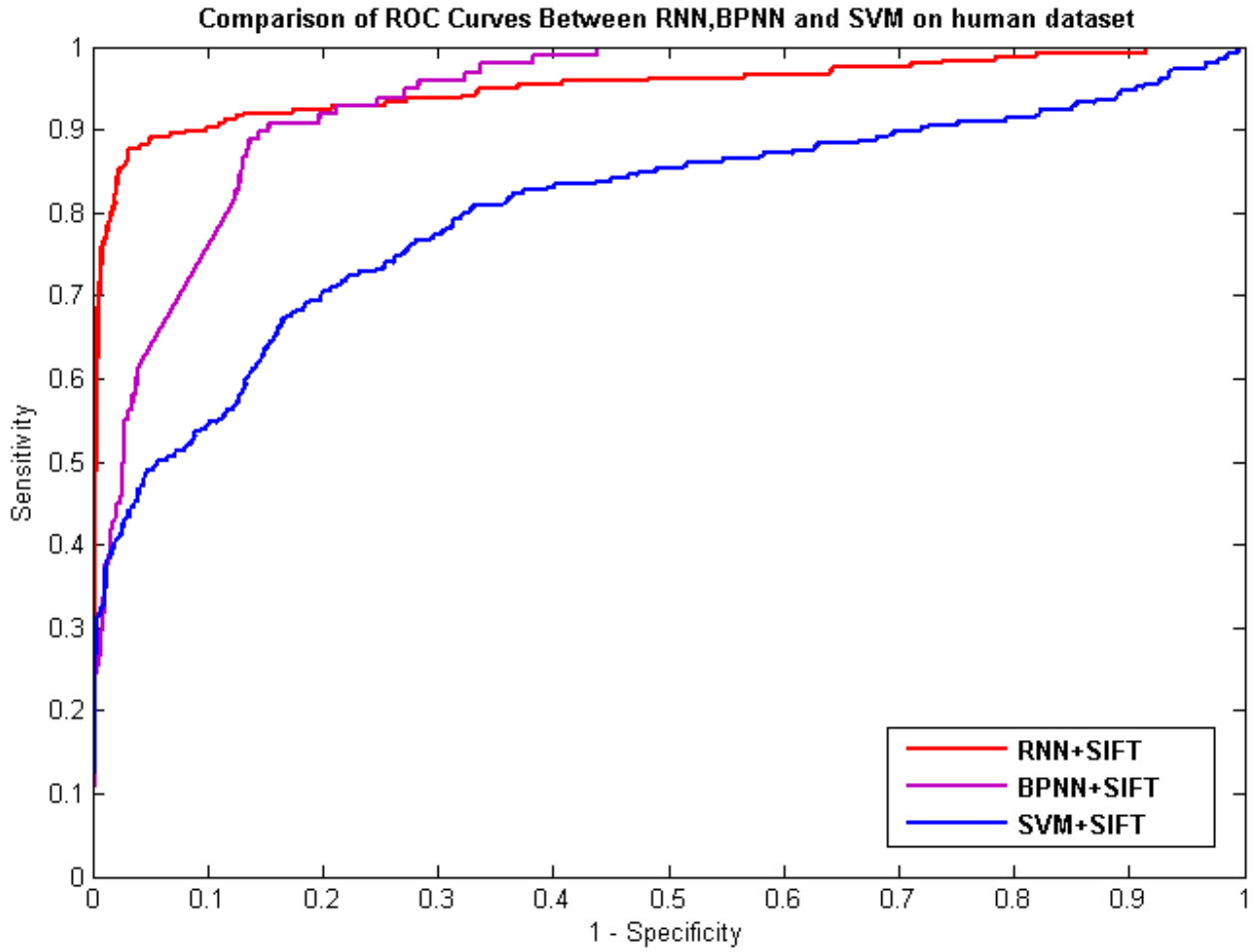

Figure 6

Comparison of ROC curves between RNN, BPNN and SVM on human dataset.