Prediction of Protein-Protein Interactions Based On Weighted Extreme Learning Machine And Speed Up Robot Features

JinXuan Zhai
Jiangsu Academy of Agricultural Sciences
Ji-Yong An (ajy@cumt.edu.cn)
School of computer science and technology https://orcid.org/0000-0001-9546-3654

Research

Keywords: PPIs, WELM, SURF, PSSM

Posted Date: August 9th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-740023/v1

Prediction of Protein-Protein Interactions Based on Weighted Extreme Learning Machine and Speed Up Robot Features

JinXuan Zhai[1], JiYong An[2,*]

[1] School of Information and Electronics Engineering, Jiangsu Vocational Institute of Architectural Technology, Xuzhou 221116, Jiangsu, China

[2]School of Computer Science and Technology, China University of Mining and Technology, Xuzhou Jiangsu 21116, China

*Corresponding author: ajy@cumt.edu.cn

Abstract:

Background：

Protein–protein interactions (PPIs) are involved in a number of cellular processes and play a key role inside cells. The prediction of PPIs is an important task towards the understanding of many bioinformatics functions and applications, such as predicting protein functions, gene-disease associations and disease-drug associations. Given that high-throughput methods are expensive and time-consuming, it is a challenging task to develop efficient and accurate computational methods for predicting PPIs .

Results：

In the study, a novel computational approach named WELM-SURF was developed to predict PPIs. The proposed method used Position Specific Scoring Matrix (PSSM) to capture protein evolutionary information and employed Speed Up Robot Features (SURF) to extract key features from PSSM of protein sequence. Weighted Extreme Learning Machine (WELM) is featured with short training time and great ability to execute classification efficiently by optimizing the loss function of weight matrix. Therefore, WELM classifier was used to carry out classification. The cross-validation results show that WELM-SURF obtains 97.36% and 95.12% of average accuracy on *yeast* and *human* dataset, respectively. The prediction ability of WELM-SURF was also compared with those of ELM-SRUF, SVM-SURF and other existing approaches. The comparison results further verify that WELM-SURF is obviously better than other methods.

Conclusion：

The experimental results proved that the WELM-SURF method is very useful for predicting PPIs and can also be applied to other bioinformatics studies of protein.

Key words: PPIs, WELM, SURF, PSSM

1. Background

As one of the most fundamental elements in living organisms, proteins make important contributions in nearly all fundamental biological processes in the cell. A large number of studies have shown that Protein-protein interactions (PPIs) play a variety of key roles in understanding functional properties of proteins and their potentials as biomarkers. More and more evidences have indicated that knowledge of PPIs can provide certain help for better understanding of the molecular mechanisms involved in biological activity, the regulation of protein function, and the underlying disease mechanisms of cellular and genetic. Although many high throughput methods, including the yeast 2-hybrid system, protein chips[1-4], and immunoprecipitation[5, 6], are typically used to identify PPIs, experimental methods for identifying PPIs are expensive and time-consuming and suffer from high rates of false positives and false negatives[7-10]. Thus, an increasing number of studies have focused on computational approaches to identify PPIs [11-13].

A number of methods are limited due to the difficulty in computing and dependence on a large number of homologous proteins [14-16]. Therefore, it is very important to identify PPIs by exploiting efficient computational approaches based on protein sequence information [17-19].

As always, a large number of researches have been conducted to predict PPIs by developing highly effective computational methods. You et. al [20]proposed a new Multi-scale Local Descriptor (MLD) feature extraction method based on protein sequence and used the Random Forest (RF) to carry out classification. The MLD can capture multi-scale local information and RF is an ensemble learning approach. Huang et. al [21] proposed a new computational method called WSRC-GE that combined weighted sparse representation (WSRC) with global coding (GE) for PPIs prediction. Wang et. al [22] presented a new computational method through combining Discrete Cosine Transform (DCT) feature extraction method with ensemble Rotation Forest (RF) classifier for PPIs prediction. An et. al [23] proposed a computational model called MKRVM-GWO that is a classification algorithm of multi-kernel RVM based on gray Wolf optimization. In order to capture the information of protein interaction, the proposed method takes full account of the local and global characteristics of protein-protein interactions position, which achieves good experimental results. Zhang et. al [24] proposed a new computational prediction model, which combined Random Tree with Genetic Algorithm to predict PPIs based on protein sequence. The prediction model obtained good prediction results. Yang et al [25] used the k-nearest neighbors to carry out classification and employed Local descriptors to extract feature from protein sequence. Guo et. al [26] presented a novel computational model called SVM-AC, which used Auto-correlation to generate feature vectors based on protein sequence and employed SVM classifier to predict PPIs. An et. al [27] proposed a new feature extraction method that can capture the continuous and discontinuous information of protein-protein interaction by using the PSSM matrix coding of local protein sequence. A number of key features can be integrated by using serial multi-feature Fusion. The above methods can explore the correlational information between protein pairs, such as, coevolution, co-localization and co-expression. [28-30]. It is highly urgent at present to develop efficient computational approaches so as to further improve the accuracy of PPIs prediction .

In the study, a novel computational method named WELM-SURF was developed to predict PPIs. The proposed method used Position Specific Scoring Matrix (PSSM) to capture protein evolutionary information and employed Speed Up Robot Features (SURF) to extract key feature from PSSM of protein sequence. The Weighted Extreme Learning Machine (WELM) is featured with short training time and great ability to efficiently execute classification by optimizing the loss function of weight matrix. Therefore, WELM classifier was used to carry out classification. The cross-validation results showed that the WELM-SURF obtained 97.36% and 95.12% of average accuracy on *yeast* and *human* dataset, respectively. The prediction ability of WELM-SURF was also compared with those of ELM-SRUF, SVM-SURF and other existing approaches. The comparison results further verify that WELM-SURF is obviously better than other methods.

2. Method

2.1. Datasets

In this study, *yeast* and *human* datasets were used as experimental datasets for evaluating the proposed method. These datasets can be obtained from the publicly available Database of Interacting Proteins (DIP)[31]. The protein sequence contained in *yeast* and *human* need to be cleaned for better executing the proposed method. The cleaning strategies are as follows: (1) The

protein sequences whose length was less than 50 residues were removed. (2) In order to eliminate bias of homologous protein sequence, the protein sequences whose length was equal or larger than 40% were considered to be homologous and thus also removed. By using the strategy, the *yeast* and *human* experimental dataset were created. In total, the *yeast* contained 5594 negative protein pairs and 5594 positive protein pairs. Similarly, the *human* contained 4262 negative protein pairs and 3899 positive protein pairs. Consequently, the experimental datasets of *yeast* and *human* are created, which contains 11188 protein pairs and 8161 protein pairs in total.

2.2. Feature Extraction Method

2.2.1 Position Specific Scoring Matrix (PSSM)

In the paper, we used Position Specific Scoring Matrix (PSSM) to extract evolution information contained protein sequence. It is because of PSSM contains not only evolution information of protein sequence, but also the position information. In the experiment, Position Specific Iterated BLAST (PSI-BLAST) tool [32] is used to transform each protein sequence into a PSSM 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 20 represent 20 amino acids, L represents the length of a given sequence, and P_{ij} represents the score of the j_{th} amino acid in the i_{th} position for the query sequence. The $P_{ij} > 0, < 0$ or $= 0$. If $P_{ij} > 0$, it indicated 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 $P_{ij} < 0$, the position is conservative and the probability of mutation is small. Smaller P_{ij} are more conservative. For obtaining highly and widely homologous sequences, the parameter e-value of PSI_BLAST tool was set to 0.001 and three iterations were selected.

2.2.2 Speed up robot features (SURF)

Speeded Up Robust Features (SURF) is the improvement of Scale Invariant Feature Transform (SIFT). In terms of algorithm execution efficiency, it runs faster than SIFT algorithm. In SIFT, Lowe uses Gaussian difference to approximate the Laplace Gaussian distribution to find the scale space. In contrast, SURF uses Box Filter to approach LoG. A major advantage of this approximation is that it is easier to compute the convolution with the box filter using the integral image, which can be done in parallel at different scales. SURF algorithm depends on the determinant and position of Hessian matrix and consists of the following two steps: feature point detection and feature proximity description..

1) Feature Point Detection

SIFT algorithm uses continuous Gaussian filters of different scales to process the image and to detect the invariant feature points through the Gaussian difference. Instead, the square filter is used in SURF to replace the Gaussian filter used in SIFT so as to achieve the Gaussian approximation. The filter can be expressed as:

$$S(x,y) = \sum_{i=0}^{x}\sum_{j=0}^{y} I(i,j)$$

The square filter can greatly improve the computation speed through using integral graph that only calculates the value the four corners of the square filter. The determinant value of hessian matrix represents the change around pixel points. Since SURF USES hessian matrix of spot detection to identify feature point whose value should be defined as the maximum or minimum value of determinant. In addition, in order to achieve scale invariance, SURF also USES the determinant of scale σ to carry out detection of feature point. For example, Given a point p=(x, y) in the graph, the Hessian matrix of scale σ is can be represented as follows:

$$H(p,\sigma) = \begin{pmatrix} L_{xx}(p,\sigma) & L_{xy}(p,\sigma) \\ L_{xy}(p,\sigma) & L_{yy}(p,\sigma) \end{pmatrix}$$

Where the $L_{xx}(p,\sigma)$, $L_{xy}(p,\sigma)$, $L_{xy}(p,\sigma)$ and $L_{yy}(p,\sigma)$ are the gray-order image after the second order differentiation. The SCALE of SURF isn't continuous Gaussian ambiguity and down sampling processing. On the contrary, it is determined by the size of square filters. The lowest scale (initial scale) of square filter of is 9×9, which is approximately σ =1.2 Gaussian filter. The size of the upper scale filter will get larger and larger, such as15×15,21×21, 27×27 ...

The transformation formula of its scale is as follows:

$$\sigma_{approx} = Currentfiltersize \times (\frac{BaseFilterscale}{BaseFilterSize})$$

2) Feature Adjacent Description.

The feature point descriptor of SURF algorithm uses the concept of Haar wavelet transform. In order to keep the feature point rotation invariant, it is necessary to give the feature point a direction. The descriptors of SURF calculate Haar wavelet transform along the X- and Y-direction within a radius dimension of 60 pixels around the feature points. The wavelet response obtained is weighted by a Gaussian function centered on the feature points. The x and y components of each wavelet response in the interval are added together to obtain a vector. Among all vectors, the longest (i.e., the one with the largest x and y components) is defined as the direction of this feature point. After the direction of the feature point is set, the pixels around it need to be based on this direction to establish the descriptor. Every 5*5 pixel points are taken as a sub-region. A total of 16 sub-regions are set within the range of 20*20 pixel points around the feature points, and the x and y directions within the sub-regions are calculated. It can be seen from the above process that since the Haar feature of 400 pixels (20×20) around each key point needs to be counted in this process, the dimensions of its descriptor vector would be 1600 (400×4) dimensions in total .Finally, a feature vector with dimensional 64 can be generated.

In the paper, we assumed that each PSSM is an image matrix. As a result, SURF feature extraction method was used to generate feature vectors and its dimensional is 64. The technology roadmap of the proposed method is shown in Figure 2.

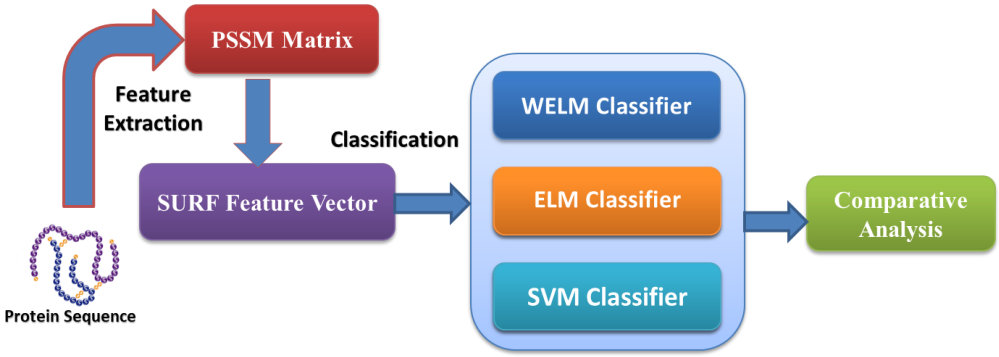

Figure 2 the technology roadmap of the proposed method

2.3 Weighted Extreme Learning Machine (WELM)

In consideration of not all samples class is evenly distributed, as a result, how to efficiently execute classification for class samples is a challenge task. Therefore, in order to solve the problem of samples classification, Weighted Extreme Learning Machine (WELM) was put forward by Zong et al [33] to solve the problem of unbalanced data classification based on Extreme Learning Machine (ELM). For the classification for PPIs datasets, we also build the WELM model based on ELM for predicting PPIs. The network structure of ELM is shown in Figure 3.

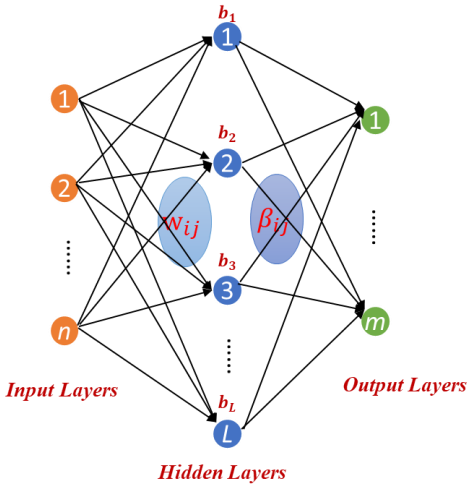

Figure 3 the network structure of ELM

Assuming there are n training samples $\{x_i, t_i\}_{i=1}^{n}$, where $x_i = \{x_{i1}, x_{i2}, x_{i3}, \ldots\ldots x_{in}\}^T \in R^n$, $t_i = \{t_{i1}, t_{i2}, t_{i3}, \ldots\ldots t_{in}\}^T \in R^m$, n represents the number of sample and m is the classification number. The output model of feedforward neural network with L hidden layer nodes can be expressed as follows:

$$\sum_{h=1}^{L} \beta_h G(a_h, b_h, x) = o_i, i = 1,2,3, \ldots\ldots, N \qquad (5)$$

Where β_h is the output weight of the h_{th} hidden layer neuron, G represents activation function of hidden layer neuron, a_h and b_h is defined as the input weight and biases of hidden layer neuron, x is input samples, o_i represents the actual output value of i_{th} training sample, t_i is the expected output of i_{th} training sample. According to the literature [15], there are N training samples $\{x_i, t_i\}_{i=1}^{n}, x_i \in R^n$. There are (a_h, b_h) andβ_h, which make $\sum_{i=1}^{L} ||o_i - t_i|| = 0$ and single-hidden layer feedforward network (SLFN) can approach the training set$\{x_i, t_i\}_{i=1}^{n}, x_i \in R^n$with zero error. The equation 1 can be simplified as follow:

$$H\beta = T \quad (6)$$

Where H and β are the output matrix and the output weight matrix of the hidden layer respectively and T is the expected output matrix corresponding training samples. The output weight of the hidden layer can be expressed as follow:

$$\hat{\beta} = \begin{Bmatrix} H^T\left(\frac{I}{C} + HH^T\right)^{-1} T, N < L \\ \left(\frac{I}{C} + H^T H\right)^{-1} H^T T, N \geq L \end{Bmatrix} \quad (7)$$

The output function of ELM can be defined as follow:

$$f(x) = h(x)\hat{\beta} = \begin{Bmatrix} h(x)H^T\left(\frac{I}{C} + HH^T\right)^{-1} T, N < L \\ h(x)\left(\frac{I}{C} + H^T H\right)^{-1} H^T T, N \geq L \end{Bmatrix} \quad (8)$$

WELM has two weighting strategies[34], one is automatic weighting and can be defined as follow:

$$w_1 = \frac{1}{Count(t_i)} \quad (9)$$

Where $\mathrm{Count}(\mathrm{t_i})$ represents the number of class t in the training sample. The other sacrifices the classification accuracy of the majority class for obtaining the classification accuracy of the minority class. This splits the minority class and the majority class into 0.618: 1(golden ratio) and is defined as follow:

$$w_2 = \begin{Bmatrix} \frac{0.618}{Count(t_i)}, t_i \in majority\ \ class \\ \frac{1}{Count(t_i)}, t_i \in minority\ \ class \end{Bmatrix} \quad (10)$$

The output weight of WELM hidden layer can be represented as follow:

$$\hat{\beta} = H^{-}T \begin{Bmatrix} H^T\left(\frac{I}{C} + WHH^T\right)^{-1} WT, N < L \\ \left(\frac{I}{C} + H^T WH\right)^{-1} H^T WT, N \geq L \end{Bmatrix} \quad (11)$$

Where the weighting matrix is a $N \times N$ diagonal matrix, and the N diagonal elements correspond to N samples. Different weights are assigned to different sample classes, and the weighting weights of the same class are the same.

This algorithm retains the advantages of simple and easy implementation of mapping function or kernel function and can be directly used for multi-category classification problems. Based on cost sensitive learning idea, WELM assigns weights to each training sample, and accordingly N weights form an N×N diagonal matrix. In general, if the training sample comes from a minority class, a relatively large weight is to be assigned to it; otherwise a relatively small weight is assigned to the training sample if it comes from a majority class. By the method of weighting, the influence of minority classes on the classification results can be enhanced and that of majority classes can be weakened correspondingly. The advantage of the WELM lies in that the link matrix ω is introduced on the basis of the extreme learning machine, which can set different weighting system for each sample that needs to be classified, so that the sample can obtain the corresponding output weight of hidden layer. Compared with BP neural network and support

vector machine, WELM has the advantages of fast training speed, simple parameter setting and strong generalization ability on the premise of guaranteeing classification performance.

The WELM has the advantage of short training time and good generalization ability and can efficiently execute classification for class samples by optimizing the loss function of weight matrix. In consideration of PPIs dataset is very class samples, so we used WELM model to predict PPIs and adopt the automatic weighting strategy in the study. The prediction flowchart of WELM-SURF model is displayed in Figure 4.

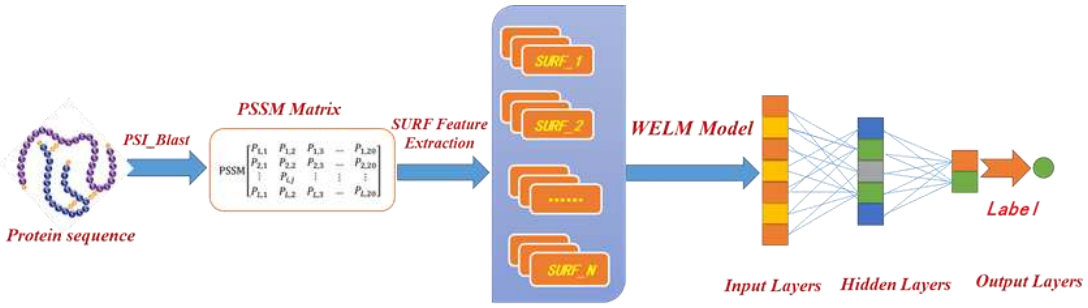

Figure 4 the prediction flowchart of WELM-SURF

2.4. Performance Evaluation

In the study, the following measures were employed to assess the performance of WELM-SURF.

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \quad (12)$$

$$TPR = \frac{TP}{TP + TN} \quad (13)$$

$$TNR = \frac{TN}{FP + TN} \quad (14)$$

$$PPV = \frac{TP}{FP + TP} \quad (15)$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}} \quad (16)$$

where Acc denotes Accuracy, TPR represents Sensitivity, TNR is specificity, PPV represents Precision, and MCC is Matthews's correlation coefficient. In addition, Receiver Operating Curve (ROC) was used to further evaluate the performance of WELM-SURF in the experiment.

3. Results and Discussion

3.1. Performance of the proposed WELM-SURF model

In this study, based on computational methods, a prediction model called WELM-SURF was proposed to predict PPIs. It used WELM to execute classification and employed SURF to generate high efficiency features. Above all, the performance of WELM-SURF was evaluated on benchmark datasets. Generally, overfitting would affect the experimental results. Therefore, the whole datasets were divided into the training datasets and independent test datasets for preventing overfitting. Specifically, the *human* dataset was randomly split into 5 equal parts, four parts of which were selected as the training set and the remaining as independent test dataset. The *yeast* dataset was also processed with the same strategy. Meanwhile, to evaluate the ability of

WELM-SURF to predict PPIs, the WELM-SURF was carried out on *yeast* and *human* dataset under five-fold cross-validation. For fair comparison, several parameters of the WELM were optimized through the grid search for ensuring fairness. To be specific, the number of Hidden layers was set to 3000, C was set to be 200, and the default values were adopted for other parameters. Table 1-2 show the results of five-fold cross-validation of WELM-SURF model on *yeast* and *human* dataset, respectively.

As can be seen from Table 1, under five-fold cross-validation ,the proposed WELM-SURF performs an average accuracy of 97.36 %, an average TPR of 96.69%, an average PPV of 97.04% and an average MCC of 92.23%. Similarly, another promising finding from Table 2 is that the WELM-SURF also achieves better prediction results on *human* dataset, whose average accuracy, average TPR, average PPR, and average MCC are 95.12%, 93.80%, 91.64% and 90.97% respectively. The prediction results demonstrate that the proposed WELM-SURF model is suitable for PPIs prediction.

The good experimental results for PPIs prediction can mainly be attributed to the SURF feature extraction method and WELM classifier. The main advantage of the WELM-SURF model is that SURF method can extract key evaluation feature from PSSM and WELM classifier has the strong classification ability of class samples. 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. In addition, it also retains plenty of prior information. This makes it possible to provide certain help in extracting the sequence evolutionary information. (2) SURF 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 to capture protein-protein interaction information and extract high efficiency features from PSSM. (3) The WELM has the advantage of short training time and good generalization ability and thus can efficiently execute classification by optimizing the loss function of weight matrix. Therefore, WELM is used to carry out classification and performs much better for identifying PPIs in the study. More specifically, the WELM can make the information of class distribution well perceived by assigning larger weight to the minority class samples and push the separating boundary from the minority class towards the majority class through using weight strategy. In this sense, it can provide an advantage for sensitive learning by assigning different weight. As is shown, the results demonstrate two things: First, SURF method is very suitable for extracting protein sequence feature; Second, the WELM classifier performs well for predicting PPIs, rendering good results.

Table 1 Fivefold cross validation results obtained by using WELM-SURF model on *yeast*

Testing set	Acc (%)	TPR (%)	PPV (%)	MCC (%)
1	97.10	96.86	97.12	92.57
2	97.41	96.31	97.21	92.80
3	96.84	95.68	97.57	91.73
4	98.51	97.28	96.33	91.70
5	96.95	97.32	96.98	92.38
Average	**97.36±0.67**	**96.69±0.69**	**97.04±0.45**	**92.23±0.50**

Table 2 Fivefold cross validation results obtained by using WELM-SURF model on *human*

Testing set	Acc (%)	TPR (%)	PPV (%)	MCC (%)

1	95.21	93.69	92.44	91.40
2	94.59	94.46	90.96	91.33
3	95.15	93.65	91.10	90.94
4	95.68	93.91	91.53	90.15
5	94.98	93.27	92.19	91.06
Average	**95.12±0.39**	**93.80±0.43**	**91.64±0.65**	**90.97±0.51**

3.2. Comparison with the ELM-based and SVM-based Methods

Experimental results demonstrate that the WELM-SURF model can accurately and efficiently predict PPIs and thus obtain good prediction results. However, to demonstrate the performance improvement of the WELM-SURF model, WELM performance was compared with that of Extreme Learning Machine (ELM) classifier and the Support Vector Machine (SVM) classifier by using the same SURF approach on *yeast* and *human* datasets, respectively. In order to ensure fair comparison, several parameter settings of ELM were optimized by means of grid search approach. Specifically, the number of hidden layers of ELM was set to 126 and other parameters would take the default value. Similarly, by using the same strategy as described above, the RBF kernel parameters of the SVM were optimized, where c was 0.3 and g was 5.2 and other parameters took the default value. In addition, SVM classifier used the LIBSVM tool [35] to carry out classification.

Table 3-6 below show the results of five-fold cross-validation of ELM-SURF and SVM-SURF on *yeast* and *human* dataset, respectively. Meanwhile, the comparison of ROC Curves on *yeast* and *human* dataset between WELM, ELM and SVM are shown in Figure 5-6 below. As outlined in Table 3-4, the ELM-SURF model achieved 94.04% average accuracy and the SVM-SURF model obtained 91.79% average accuracy on *yeast* dataset. Similarly, as can be seen from Table 5-6, the results of average accuracy 92.04% and 89.58% are obtained by the ELM-SURF model and the SVM-SURF model on *human* dataset, respectively. When comparing the results with those of ELM-SURF and SVM-SURF, it must be pointed out that the performance of WELM classifier is significantly better than the other two classifiers. At the same time, from Figure 5 and Figure 6, the ROC curves of WELM classifier are also significantly better than that of the other two classifiers. A major reason for good prediction results is that the WELM has the advantage of short training time and good generalization ability and can efficiently execute classification for class samples by optimizing the loss function of weight matrix. Specifically, it can make the information of class distribution well perceived by assigning larger weight to the minority class samples and pushing the separating boundary from the minority class towards the majority class by means of weight strategy. From the above analysis, this paper comes to the conclusion that the proposed WELM-SURF model is useful tools for PPIs prediction, as well as other bioinformatics tasks.

Table 3 Fivefold cross validation results obtained by using ELM-SURF model on *yeast*

Testing set	Acc (%)	TPR (%)	PPV (%)	MCC (%)
1	94.05	89.14	90.89	85.82
2	95.07	90.83	90.74	84.09
3	93.02	86.02	91.19	84.95
4	93.87	87.26	90.42	82.95

5	94.17	88.19	91.26	84.27
Average	**94.04±0.73**	**88.34±1.87**	**90.90±0.34**	**84.42±1.06**

Table 4 Fivefold cross validation results obtained by using SVM-SURF model on *yeast*

Testing set	Acc (%)	TPR (%)	PPV (%)	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**

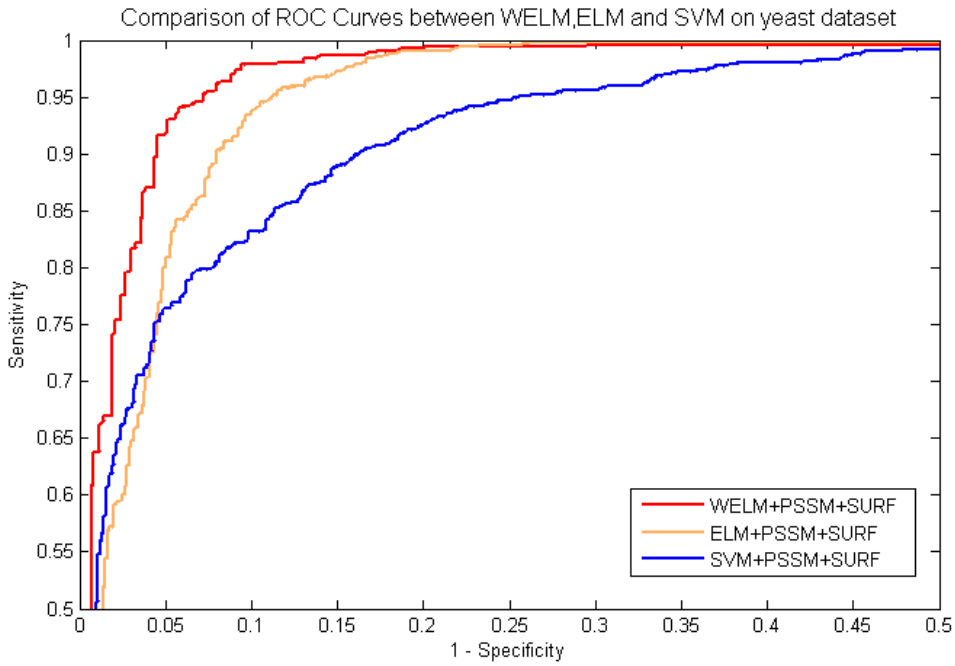

Figure 5 Comparison of ROC curves between WELM, ELM and SVM on *yeast* dataset.

Table 5 Fivefold cross validation results obtained by using ELM-SURF model on *human*

Testing set	Acc (%)	TPR (%)	PPV (%)	MCC (%)
1	92.98	87.85	87.73	81.24
2	91.89	93.23	86.31	83.98
3	91.31	85.68	87.59	78.48
4	92.28	87.89	86.72	80.07
5	91.76	85.49	86.18	77.61

Average	**92.04±0.63**	**88.03±3.13**	**86.91±0.72**	**80.28±2.50**

Table 6 Fivefold cross validation results obtained by using SVM-SURF model on *human*

Testing set	Acc (%)	TPR (%)	PPV (%)	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**

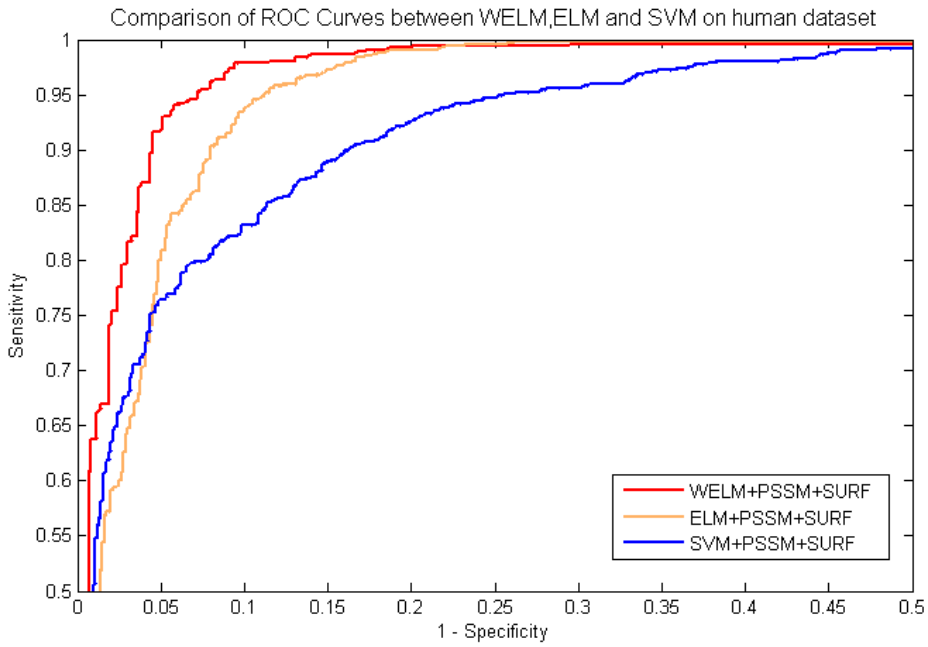

Figure 6 Comparison of ROC curves between WELM, ELM and SVM on *human* dataset.

3.3. Comparison with Other Methods

To further validate the prediction ability of the WELM-SURF model, the comparison results between WELM-SURF model and the previous methods on *yeast* and *human* dataset were displayed in Table7-8. As can be seen from Table 7, the average accuracy of WELM-SURF is obviously higher than those of the other six approaches on *yeas*t dataset. Similarity, Table 8 displays the prediction accuracy obtained by WELM-SURF model is also significantly better than those of the other six methods on *human* dataset. A similar conclusion could be reached by comparing the results from Table 7-8 that the proposed WELM-SURF model has an excellent prediction capability and can be used for the quality prediction of PPIs.

Table 7 The comparison between WELM-SURF and other methods on *yeast* dataset

Model	Acc (%)	TNR (%)	TPR (%)	MCC
Guo[36]	89.33±2.67	89.93±3.60	88.77±6.16	*N/A*
Zhou[37]	88.56±0.33	87.37±0.22	89.50±0.60	77.15±0.68
Yang[38]	86.15±1.17	81.03±1.74	90.24±1.34	*N/A*
Wong[39]	93.92±0.36	91.10±0.31	96.45±0.45	88.86±0.63
Huang[40]	96.28±0.52	92.64±1.00	99.92±0.32	92.82±0.97
Our method	**97.36±0.67**	**96.69±0.69**	**97.04±0.45**	**92.23±0.50**

Table 8 The comparison between WELM-SURF and other methods on *human* dataset

Model	Acc (%)	TNR (%)	TPR (%)	MCC
Nanni[41]	83.00	86.00	85.10	*N/A*
Nanni[42]	84.00	86.00	84.00	*N/A*
Lumin[43]	86.60	86.70	85.00	*N/A*
You [44]	92.83	89.32	96.13	86.85
Nanni[45]	93.90	*N/A*	*N/A*	*N/A*
Our method	**95.12**	**93.80**	**91.64**	**90.97**

4. Conclusion

In the study, a new computational method called WELM-SURF was put forward for PPIs prediction, which combines the Weighted Extreme Learning Machine (WELM) with Speed up robot features (SURF) to predict PPIs based on protein evolutionary information. By comparing with experimental results, the performance of WELM-SURF is significantly better than those of the ELM, SVM and other previous methods in the domain. The excellent performance of WELM-SURF mainly attributes to the following several important factors: (1) PSSM contains not only the position information, but also the evolution information of protein sequence. In addition, it also retains plenty of prior information. This makes it possible to provide certain help in extracting the sequence evolutionary information. (2) SURF 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 to capture protein-protein interaction information and extract high efficiency features from PSSM. (3) The WELM has the advantage of short training time and good generalization ability and can efficiently execute classification by optimizing the loss function of weight matrix. The WELM can make the information of class distribution well perceived by assigning larger weight to the minority class samples; meanwhile it pushes the separating boundary from the minority class towards the

majority class by means of weight strategy and can provide an advantage for sensitive learning by assigning different weights. Therefore, it can be drawn that the proposed WELM-SURF model is useful and can execute incredibly well for PPIs prediction as well as other bioinformatics tasks.

Abbreviations:

PPIs: Protein-protein interactions
WELM: Weighted Extreme Learning Machine
SURF: Speed up robot features
PSSM: Position Specific Scoring Matrix
SVM: Support Vector Machine
ELM: Extreme Learning Machine
PSI-BLAST: Position-Specific Iterated BLAST
Acc: Accuracy
TNR: True Negative Rate
TPR: True Positive Rate
MCC: Matthews Correlation Coefficient
PPV: Positive Predictive Value
ROC: Receiver Operating Curve

Declarations

Ethics approval and consent to participate： Not applicable
Consent for publication： Not applicable
Availability of data and material： These datasets can be obtained from the publicly available Database of Interacting Proteins (DIP) [31]
Competing interests: The authors declare no conflict of interest.
Competing interests: The authors declare no potential conflicts of interest with respect to the research, authorship, and publication of this article.
Funding: This work was supported by 'the Fundamental Research Funds for the Central Universities, No. 2019XKQYMS88." and 'Project of Construction Science and Technology of Jiangsu, No.2018ZD084".The funder had no role in study design, data collection and preparation of the manuscript.
Author Contributions: ZJX and AJY conceived the algorithm, carried out analyses, prepared the data sets, carried out experiments, and wrote the manuscript.
Acknowledgments: The authors would like to thank all the guest editors and anonymous reviewers for their constructive advice.

References

1. Palopoli N, Edwards R: **Large-scale prediction of short linear motifs using structural information from protein-protein interactions**. *Bmc Bioinformatics* 2015.
2. Foltman M, Sanchez-Diaz A: **Studying Protein–Protein Interactions in Budding Yeast Using Co-immunoprecipitation**. *Methods in Molecular Biology* 2016, **1369**:239.
3. Zhu H, Bilgin M, Bangham R, Hall D, Casamayor A, Bertone P, Lan N, Jansen R, Bidlingmaier S, Houfek T *et al*: **Global analysis of protein activities using proteome chips**. *Science* 2001, **293**(5537):2101-2105.

4. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K *et al*: **Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry**. *Nature* 2002, **415**(6868):180-183.
5. Schwikowski B, ., Uetz P, ., Fields S, . **A network of protein-protein interactions in yeast**. *Nature Biotechnology* 2000, **18**(12):1257.
6. Rain JC, Selig L, ., Reuse H, De, Battaglia V, ., Reverdy C, ., Simon S, ., Lenzen G, ., Petel F, ., Wojcik J, ., Sch?Chter V, . **The protein-protein interaction map of Helicobacter pylori**. *Nature* 2001, **409**(6817):211-215.
7. An JY, Zhou Y, Zhang L, Niu Q, Wang DF: **Improving Self-interacting Proteins Prediction Accuracy Using Protein Evolutionary Information and Weighed-Extreme Learning Machine**. *Current Bioinformatics* 2019.
8. Jia J, Xiao X, Liu B: **Prediction of Protein-Protein Interactions with Physicochemical Descriptors and Wavelet Transform via Random Forests**. *J Lab Autom* 2015, **21**(3):368-377.
9. Smits AH, Vermeulen M: **Characterizing Protein-Protein Interactions Using Mass Spectrometry: Challenges and Opportunities**. *Trends in Biotechnology* 2016, **34**(10):825-834.
10. Gao, Lin, Wang, Bingbo, Deng: **ppiPre: predicting protein-protein interactions by combining;heterogeneous features**. *Bmc Systems Biology* 2013, **7**(Suppl 2):S8-S8.
11. Hue M, Riffle M, Vert JP, Noble WS: **Large-scale prediction of protein-protein interactions from structures**. *Bmc Bioinformatics* 2010, **11**(1):1-9.
12. Lu L, Lu H, J: **MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading**. *Proteins-structure Function & Bioinformatics* 2010, **49**(3):350-364.
13. Yu C, Dong X: **Computational analyses of high-throughput protein-protein interaction data**. *Current Protein & Peptide Science* 2003, **4**(3):-.
14. An JY, Zhang L, Zhou Y, Zhao YJ, Wang DF: **Computational methods using weighed-extreme learning machine to predict protein self-interactions with protein evolutionary information**. *Journal of Cheminformatics* 2017, **9**(1):47.
15. Keskin O, Tuncbag N, Gursoy A: **Predicting Protein-Protein Interactions from the Molecular to the Proteome Level**. *Chemical Reviews* 2016, **116**(8):4884.
16. You ZH, Li X, Chan KC: **An improved sequence-based prediction protocol for protein-protein interactions using amino acids substitution matrix and rotation forest ensemble classifiers**. *Neurocomputing* 2016, **228**:277-282.
17. 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**. *Genomics* 2014, **104**(6):496-503.
18. Zhang, Shengli: **Accurate prediction of protein structural classes by incorporating PSSS and PSSM into Chou\"s general PseAAC**. *Chemometrics & Intelligent Laboratory Systems* 2015, **142**:28-35.
19. 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.
20. You ZH, C. CKC, Pengwei H, Franca F: **Predicting Protein-Protein Interactions from Primary Protein Sequences Using a Novel Multi-Scale Local Feature Representation Scheme and the Random Forest**. *Plos One*, **10**(5):e0125811-.

21. 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.
22. Wang L, You Z-H, Xia S-X, Liu F, Chen X, Yan X, Zhou Y: **Advancing the prediction accuracy of protein-protein interactions by utilizing evolutionary information from position-specific scoring matrix and ensemble classifier**. *Journal of Theoretical Biology*, **418**(Complete):105-110.
23. 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.
24. Lei Z: **Sequence-Based Prediction of Protein-Protein Interactions Using Random Tree and Genetic Algorithm**. In: *International Conference on Intelligent Computing: 2012*.
25. Yang L, Xia JF, Gui J: **Prediction of protein-protein interactions from protein sequence using local descriptors**. *Protein & Peptide Letters* 2010, **17**(9):1085.
26. 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.
27. 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.
28. 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.
29. 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.
30. 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 Training Datasets**. *Molecules* 2015, **21**(1):E95.
31. Xenarios I, Salwinski L, Duan XQJ, Higney P, Kim SM, Eisenberg D: **DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions**. *Nucleic Acids Research* 2002, **30**(1):303-305.
32. 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.
33. Zong WW, Huang GB, Chen YQ: **Weighted extreme learning machine for imbalance learning**. *Neurocomputing* 2013, **101**:229-242.
34. Pan WT: **A new Fruit Fly Optimization Algorithm: Taking the financial distress model as an example**. *Knowledge-Based Systems*, **26**:p.69-74.
35. Chih-Chung, Chang, Chih-Jen, Lin: **LIBSVM: A library for support vector machines**.
36. Yanzhi G, Lezheng Y, Zhining W, Menglong L: **Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences**. *Nucleic Acids Research* 2008, **36**(9):3025-3030.
37. Yu ZZ, Yun G, Ying YZ: **Prediction of Protein-Protein Interactions Using Local Description of Amino Acid Sequence**. *Communications in Computer & Information Science* 2011,

202:254-262.

38. Lei Y, Jun-Feng X, Jie G: **Prediction of protein-protein interactions from protein sequence using local descriptors**. *Protein & Peptide Letters* 2010, **17**(9):-.
39. Wong L, You ZH, Li S, Huang YA, Liu G: **Detection of Protein-Protein Interactions from Amino Acid Sequences Using a Rotation Forest Model with a Novel PR-LPQ Descriptor**. 2015.
40. Yu-An H, Zhu-Hong Y, Xin G, Leon W, Lirong W: **Using Weighted Sparse Representation Model Combined with Discrete Cosine Transformation to Predict Protein-Protein Interactions from Protein Sequence**. *Biomed Research International* 2015, **2015**:902198.
41. Nanni L: **Letters: Fusion of classifiers for predicting protein-protein interactions**: Elsevier Science Publishers B. V.; 2005.
42. Nanni L: **Letters: Hyperplanes for predicting protein-protein interactions**: Elsevier Science Publishers B. V.; 2005.
43. Nanni L, Lumini A: **An ensemble of K-local hyperplanes for predicting protein-protein interactions**. *Bioinformatics* 2006, **22**(10):1207-1210.
44. Yu-An, Huang, Zhu-Hong, You, Xing, Chen, Keith, Chan, Xin, Luo: **Sequence-based prediction of protein-protein interactions using weighted sparse representation model combined with global encoding**. *BMC bioinformatics* 2016.
45. Loris N, Alessandra L, Sheryl B: **An Empirical Study of Different Approaches for Protein Classification**. *Scientificworldjournal* 2014, **2014**:236717.