An Efficient Computational Model for Class Imbalance Problem in Self-Interaction Proteins Prediction

Ji-Yong An (ajy@cumt.edu.cn)
China University of Mining and Technology https://orcid.org/0000-0001-9546-3654
Yong Zhou
China University of Mining and Technology
Zi-Ji Yan
China University of Mining and Technology
Yu-Jun Zhao
China University of Mining and Technology

Research article

Keywords: SIPs, WELM, SURF, PSSM

Posted Date: July 9th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-36603/v1

An efficient computational model for class imbalance problem in Self-Interaction Proteins Prediction

Ji-Yong An[1,2], Yong Zhou[1,2], Zi-Ji Yan[1], Yu-Jun Zhao[1]

(ajy@cumt.edu.cn, ajysjm@163.com, yanzj@cumt.edu.cn, zyj@cumt.edu.cn)

[1]Engineering Research Center of Mine Digitalization (China University of Mining and Technology)

Ministry of Education, China

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

Xuzhou Jiangsu 21116, China

Corresponding author: Ji-Yong An

Abstract:

Background：

Self-interaction Proteins (SIPs) play a key role in a variety of biological activities of organisms. In consideration of the time-consuming and expensive of high-throughput methods, and the number of positive and negative samples is very imbalanced in SIPs datasets. How to develop accurate and efficient computational approaches for assisting and accelerating the study of identifying SIPs is a challenging task.

Results：

In the work, we proposed a new computational method called WELM-SURF for predicting SIPs. More specifically, for exploiting protein sequence feature, Position Specific Scoring Matrix (PSSM) is applied to capturing protein evolutionary information and Speed up robot features (SURF) is employed to extract key feature of protein sequence from PSSM. Take account of the advantage that the Weighted Extreme Learning Machine (WELM) has short training time, good generalization ability, and most importantly ability to efficiently execute classification for imbalanced class samples by optimizing the loss function of weight matrix. Therefore, the WELM classifier is used to perform classification based on extracted features for predicting SIPs. A large number of experiments show that the average accuracy of WELM-SURF is 95.25% and 98.79% on *yeast* and *human* dataset, respectively. We also compared our performance with Extreme Learning Machine (ELM), the state-of-the-art Support Vector Machine (SVM), and other existing methods. Compared with the experimental results, the performance of WELM-SURF in this domain is obviously better than ELM, SVM and other previous methods.

Conclusion：

These experimental results proved that the proposed WELM-SURF model is competent for predicting SIPs with high accuracy and robustness. It is anticipated that the WELM-SURF method is a useful computational tool to facilitate widely bioinformatics studies related to SIPs prediction. For further encouraging future proteomics research, we developed a freely available web server called WELM-SURF-SIPs. It is available at http://219.219.62.123:8888/WELMSURF/ and includes SIPs datasets and source code.

Key words: SIPs, WELM, SURF, PSSM

1. Background

A large number of studies have shown that Protein-protein interactions (PPIs) play a variety of key roles in many important biological activities. However, whether proteins can interact with their partners is an important research direction of proteomics research. Self-interactions protein (SIPs) refers to two or more copies of a protein that is the same copies and is represented by the same gene, which can interact with each other and is considered as a special type of PPIs. This

might bring about the formation of homo-oligomer problem. In recent years, many studies have proved that SIPs plays an important role in the evolution of various cellular physiological functions and protein-protein interaction networks (PPINs) [1-3]. Therefore, it is important for a protein to express function through its own interactions. The research related to SIPs can provide a 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. Homologous oligomerization is an important function of biological activity and plays an absolutely important role in gene expression regulation, signal transduction, immune response and enzyme activation [4-8]. In addition, many previous studies have revealed that the diversity function of proteins can be different degrees expanded through SIPs without increasing genome length. SIPs can also improve the stability and prevent the denaturation of proteins through reducing their surface area[9, 10]. As a result, it is increasingly important for developing reliable and efficient computational methods to predict SIPs based on protein sequences.

As always, a large number of researches have been devoted to develop reliable and highly effective computational approaches to predict PPIs. You et al [11]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 [12] proposed a new computational method called WSRC-GE that combined weighted sparse representation (WSRC) with global coding (GE) for predicting PPIs. Wang et al [13] presented a new computational method through combining Discrete Cosine Transform (DCT) feature extraction method with ensemble Rotation Forest (RF) classifier for predicting PPIs. An et al [14] 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 characteristics of local and global of protein-protein interactions position, which achieves good experimental results. Zhang et al [15] 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 [16] used the k-nearest neighbors for carrying out classification and employed Local descriptors to extract feature from protein sequence. Guo et al [17] presented a novel computational model called SVM-AC, which used Autocorrelation to generate feature vectors based on protein sequence and employed SVM classifier to predict PPIs. An et al [18] proposed a new feature extraction method that can capture protein-protein interaction information of continuous and discontinuous 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. However, this information is not sufficient to predict SIP. In addition, the PPIs dataset does not contain PPIs between the same protein partners and SIPs dataset is very imbalanced. In the previous study, Liu et al [1] proposed a prediction model called SLIPPER for predicting SIPs, which integrate multi representative known properties. As far as we know, many research results have been reported about SIPs in recent studies [19-21]. However, these methods have an obvious disadvantage that cannot deal with the proteins without covered current human interatomic and solve the class imbalance problem in SIPs. For these reasons, it is an urgent work at present for developing efficient computational approaches for solving the imbalanced class

classification of predicting SIPs.

In the paper, we proposed a new computational method called WELM-SURF for predicting SIPs. More specifically, for exploiting protein sequence feature, Position Specific Scoring Matrix (PSSM) is applied to capturing protein evolutionary information and Speed up robot features (SURF) is employed to extract key feature of protein sequence from PSSM. Take account of the advantage that the Weighted Extreme Learning Machine (WELM) has short training time, good generalization ability, and most importantly ability to efficiently execute classification for imbalanced class samples by optimizing the loss function of weight matrix. Therefore, the WELM classifier is used to perform classification based on extracted features for predicting SIPs. A large number of experiments show that the average accuracy of WELM-SURF is 95.25% and 98.79% on yeast and human dataset, respectively. We also compared our performance with Extreme Learning Machine (ELM), the state-of-the-art Support Vector Machine (SVM), and other existing methods. Compared with the experimental results, the performance of WELM-SURF in the domain is obviously better than ELM, SVM and other previous methods. These experimental results proved that the proposed WELM-SURF model is competent for predicting SIPs with high accuracy and robustness. It is anticipated that the WELM-SURF method is a useful computational tool to facilitate widely bioinformatics studies related to SIPs prediction.

2. Method

2.1. Datasets

The PPIs datasets from the previous research, including DIP [22], BioGRID[23], IntAct[24], InnateDB [25] and MatrixDB [26] and the Uniprot database contains 20,199 curated *human* protein sequences [27]. In order to construct the SIPs experimental dataset, the protein sequences that only interact with themselves were selected from the above PPIs dataset and the type of interaction has been defined as "direct interaction" in the relevant database. In order to assess the performance of WELM-SURF, 2994 human self-interaction protein sequences were screened in for creating the experimental dataset by adopting 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) to create the positive samples, one of the following conditions must be satisfied: (a) At least two kinds of large scale experiments or one small-scale experiment has detected its Self-interactions; (b) the Uniprot dataset has defined the protein sequences as homopolymer; (c) it has been reported by at least two publications for its Self-interactions;(3) to construct negative samples, we removed all types of SIPs from the entire human proteome (including proteins annotated as "direct interactions" and more broadly as "physical associations") and the Uniprot database. Finally, 15,938 non-SIPs were selected as negatives samples and 1441 SIPs were selected as positives samples to construct the human dataset[28]. At the same time, we also construct the *yeast* dataset, which includes 5511 negative samples and 710 positive samples by using the same strategy [28]. There are about 8 times as many positive samples as negative samples for yeast dataset and about 11 times for human dataset. Therefore, SIPs datasets are very imbalanced class samples.

2.2. Feature Extraction Method

2.2.1 Position Specific Scoring Matrix (PSSM)

Due to proteins are functionally conserved, the prediction performance can be improved by using the evolutionary information of protein sequence. The position-specific scoring matrix

(PSSM) contains not only the position information of the protein sequence, but also the evolution information that reflects the conservative function of protein. In the experiment, each protein sequence was converted a $L \times 20$ PSSM by using Position Specific Iterated BLAST (PSI-BLAST) tool [29],where L represents the length of different protein sequences. Therefore, we employed the PSSM for extracting the sequence evolutionary information because of its advantage in the paper. The diagram of PSSM is displayed in Figure 1.

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

Figure 1 the diagram of PSSM

Where 20 are 20 different amino acids, P_{ij} represent the probability that the i_{th} amino acid in the sequence is mutated to the j_{th} type amino acid during biological evolution. The P_{ij} is greater than 0, equal to 0 and less than 0. If the P_{ij} is a positive number that indicates the i_{th} amino acid can be easily mutated to the $\mathrm{j_{th}}$ amino acid. In practice, the larger number of P_{ij} means a higher mutation probability. Conversely, if P_{ij} is negative number, it means the mutation probability is small, and a smaller P_{ij} number indicates more conservative. For using evolutionary information of protein sequences to capture more key features, we converted each SIP's sequence into a PSSM through employing PSI-BLAST tool. In the experiment, we set the parameter of PSI_BLAST's e-value is 0.001 and selected three iterations for obtaining widely and highly homologous sequences.

2.2.2 Speed up robot features (SURF)

Speed up robot features (SURF)[30] feature extraction algorithm is an improvement of Scale Invariant Feature Transform (SIFT) algorithm[31, 32], which runs faster than SIFT in algorithm execution efficiency. The SIFT uses Gaussian differences to approximate Laplace Gauss distribution to find scale space. However, the SURF uses Box Filter to approximate LOG. The major advantage of SURF is that it is easier to calculate the convolution with the box filter by using the integrated image, which can be done in parallel at different scales. The execution of the SURF algorithm depends on the determinant of the Hessian matrix and the determinant of the position. The SURF algorithm includes the following two steps: feature point detection and feature adjacent description.

1) Feature Point Detection

The SURF uses continuous Gaussian filters of different scales to process image and detects feature points of mesoscale invariant through Gaussian differences. SURF can represent Gaussian fuzzy approximation by using the square filter to replace the Gaussian filters of SIFT. 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 descriptor of SURF uses the concept of Hal wavelet transform. In order to ensure the rotation invariance of feature point, each feature point is assigned a direction. The SURF descriptors calculate the Hal wavelet transform of 6σ pixels of direction of X and Y around feature point. A vector can be obtained by add components of corresponding X and Y of the wavelet in each interval. The longest (the largest X and Y components) of all vectors is the direction of the feature point. After the direction of the feature point is selected, the descriptor of feature point can be created by using the direction of surrounding pixels. For example, the 5×5 pixel points were defined as a sub region. As a result, a number of 16 sub regions can be generated by extracting the range of 20*20 pixel points around the feature point and the $\sum dx$ and $\sum dy$ of the Hal wavelet transform in the X and Y directions within the sub region can be calculated. Finally, a feature vector with dimensional 64 can be generated.

In the experiment, we used SURF method to create feature vectors whose dimensional is 64. Figure 2 shows the flow diagram of our method.

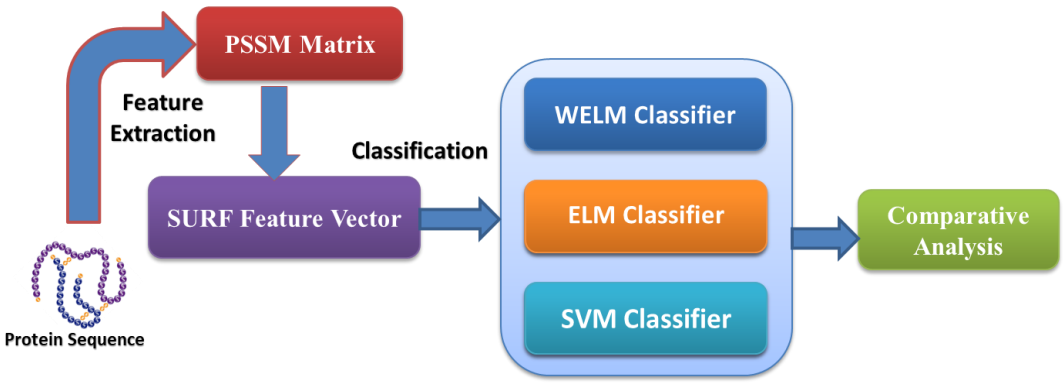

Figure 2 the flow diagram of our method

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 imbalanced class samples is a challenge task. Therefore, in order to solve the problem of imbalanced samples classification, Zong et al [33] proposed a Weighted Extreme Learning Machine (WELM) based on Extreme Learning Machine (ELM). For the classification for imbalanced SIPs datasets, we also build the WELM model based on ELM for predicting SIPs.

The network structure of ELM is as follows:

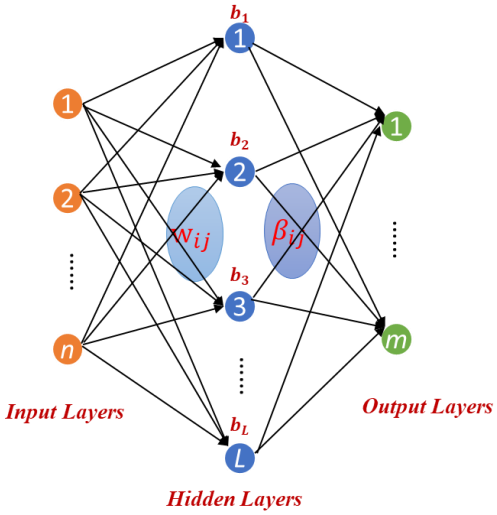

Figure 3 the network structure of ELM

Assuming there are n training samples $\{x_i, t_i\}_{i=1}^{n}$, where $x_i = \{x_{i1}, x_{i2}, x_{i3}, \dots \dots x_{in}\}^T \in R^n$, $t_i = \{t_{i1}, t_{i2}, t_{i3}, \dots \dots t_{in}\}^T \in R^m$, 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:

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

Where β_h is the output weight of the $\mathbf{h}_{th}$ 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 $Count(t_i)$ 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.

The WELM has the advantage of short training time and good generalization ability and can efficiently execute classification for imbalanced class samples by optimizing the loss function of weight matrix. Considering that SIPs dataset is very imbalanced class samples and advantage of WELM model in imbalance classification. As a result, the WELM classifier was used to predict SIPs by employing the automatic weighting strategy. The prediction flow diagram of WELM-SURF model is shown in Figure 4.

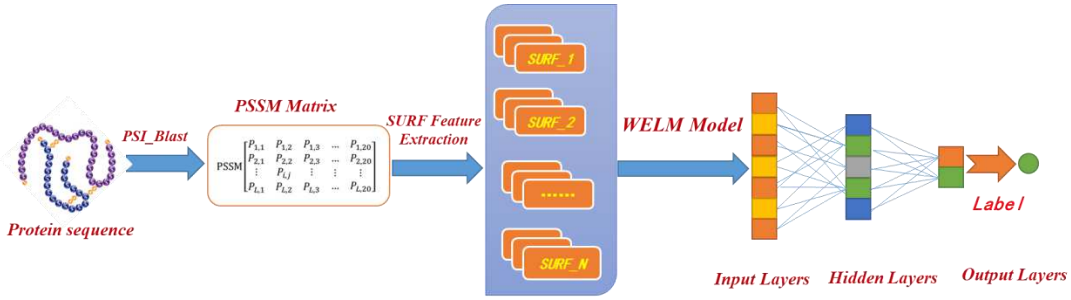

Figure 4 the prediction flow diagram of WELM-SURF

2.4. Performance Evaluation

The following measures were used to evaluate the prediction performance of WELM-SURF in the work.

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

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

$$\text{PPV} = \frac{TP}{FP + TP} \quad (14)$$

$$\text{MCC} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}} \quad (15)$$

Where Acc represents Accuracy, TPR is Sensitivity, PPV is Precision and MCC represents Matthews's correlation coefficient. TP and TN represent the count of real interaction and real

non-interaction protein sequence pairs correctly predicted. FP and FN is the number of real non-interaction and real interaction protein sequence pairs mistakenly predicted. Meanwhile, Receiver Operating Curve (ROC) was employed to further assess the prediction performance of WELM-SURF in the work.

3. Results and Discussion

3.1. Performance of the proposed WELM-SURF model

In this work, we proposed a prediction model based on computational method to predict SIPs, called WELM-SURF, which used WELM to execute imbalanced classification and employed SURF to generate high efficiency features. Above all, the performance of WELM-SURF was evaluated on benchmark datasets. The overfitting usually affects the prediction results. As a result, in order to prevent overfitting, the whole dataset is divided into training dataset and independent test dataset. In other words, we randomly divided the human dataset into 5 equal parts, of which 4 parts were used as training dataset and the rest as independent test dataset. The same strategy was also applied to the *yeast* dataset. At the same time, to evaluate WELM-SURF's ability of predicting SIPs, the WELM-SURF is carried out on *yeast* and *human* dataset under five-fold cross-validation. In order to ensure the fairness of comparison, several parameters of the WELM classifier were optimized by grid search algorithm. Where the number of Hidden layers is 3000, C = 200 and other parameters were set up the default value. Table 1-2 shows 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 95.25 %, an average TPR of 93.05%, an average PPV of 94.35% and an average MCC of 86.44% As shown in Table 2, the WELM-SURF model also obtained very good experimental results on *human* dataset, whose average accuracy, average TPR, average PPR, and average MCC are 98.79%, 95.15%, 96.65% and 91.89% respectively. The prediction results demonstrated that our WELM-SURF is suitable for SIPs prediction.

The WELM-SURF can obtain very good prediction results, this attributes to SURF can capture key features from PSSM and WELM classifier has the strong classification ability for imbalanced class samples. Specifically, there are three main reasons: (1) The PSSM contains not only the position information of the protein sequence, but also the evolution information that reflects the conservative function of protein and a number of prior information. Therefore, it can provide a certain help in extracting evolutionary information of protein sequence and capture key SIP features. (2) SURF can improve computational speed compared to SIFT. The main advantage of SURF that it 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 can capture self-protein interaction information and extract high efficiency features from PSSM. (3) For the sake of SIPs datasets are very imbalanced class samples and the WELM has the advantage of short training time and good generalization ability and can efficiently execute classification for imbalanced class samples by optimizing the loss function of weight matrix. Therefore, WELM is use to carry out classification and performs much better for identifying SIPs in the study. More specifically, the WELM can better perceive the distribution information of imbalanced class 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. This makes it can provide help in sensitive learning by assigning different weight. As a result, the results demonstrate two things. First, SURF feature extraction

approach is suitable for extracting SIP feature from the PSSM of protein sequence, and secondly, the WELM classifier can obtain good prediction results for predicting SIPs by imbalanced learning.

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

Testing set	Acc (%)	TPR (%)	PPV (%)	MCC (%)
1	93.65	91.90	95.22	86.09
2	94.19	90.93	94.88	86.93
3	93.52	91.02	94.32	85.16
4	93.23	91.53	93.79	87.12
5	94.01	91.58	93.67	86.90
Average	**95.25±0.38**	**93.05±3.51**	**94.35±0.67**	**86.44±0.82**

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

Testing set	Acc (%)	TPR (%)	PPV (%)	MCC (%)
1	97.85	95.32	96.38	91.80
2	98.18	95.53	97.35	93.01
3	98.17	96.06	96.32	92.40
4	97.66	94.70	96.81	91.63
5	97.13	94.14	96.38	90.62
Average	**98.79±0.43**	**95.15±0.75**	**96.65±0.44**	**91.89±0.89**

3.2. Comparison WELM-SURF method with the ELM-based and SVM-based

Experimental results demonstrate that the WELM-SURF model can accurately and efficiently predict SIPs and obtain better experimental results. However, to demonstrate the performance improvement of WELM-SURF model, the performance of WELM classifier was compared with the performance of ELM classifier and the SVM classifier through employing the same SURF feature extraction method on *yeast* and *human* datasets, respectively. For fair comparison, several parameter of ELM were optimized through employing the same grid search method. More specifically, the number of hidden layers of ELM is set to 126 and other parameters take the default value. At the same time, the RBF kernel parameters of the SVM were optimized by using the same strategy, where c = 0.3 and g = 5.2 and other parameters were set up the default value. In the experiment, LIBSVM tool [35] was used to execute classification.

Table 3-6 displays the prediction results of five-fold cross-validation of ELM-SURF and SVM-SURF on *yeast* and *human* dataset, respectively. At the same time, the comparison of ROC Curves between WELM, ELM and SVM on *yeast* and *human* dataset are shown in Figure 5-6. As can be seen from Table 3-4, the ELM-SURF model and the SVM-SURF obtain average accuracy of 92.04% and 89.58% on *yeast* dataset, respectively. Similarly, as outlined in table 5-6, the ELM-SURF obtained 94.04% average accuracy and the SVM-SURF achieved 91.79% average accuracy on *human* dataset. It should be emphasized that the classification ability of WELM is obviously better the other classifiers by comparing these experimental results. Meanwhile, as can be seen from Figure 5 and Figure 6, the ROC curves of WELM are also significantly better than the other classifiers. One important reason is that the WELM focus on the imbalanced class classification relative to ELM and SVM. It has the advantage of short training time and good

generalization ability and can efficiently execute classification for imbalanced class samples by optimizing the loss function of weight matrix. Specifically, the WELM can better perceive the distribution information of imbalanced class 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. From the above analysis, the paper comes to the conclusion that the proposed WELM-SURF model is a useful tool for predicting SIPs, as well as other bioinformatics tasks.

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

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

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**

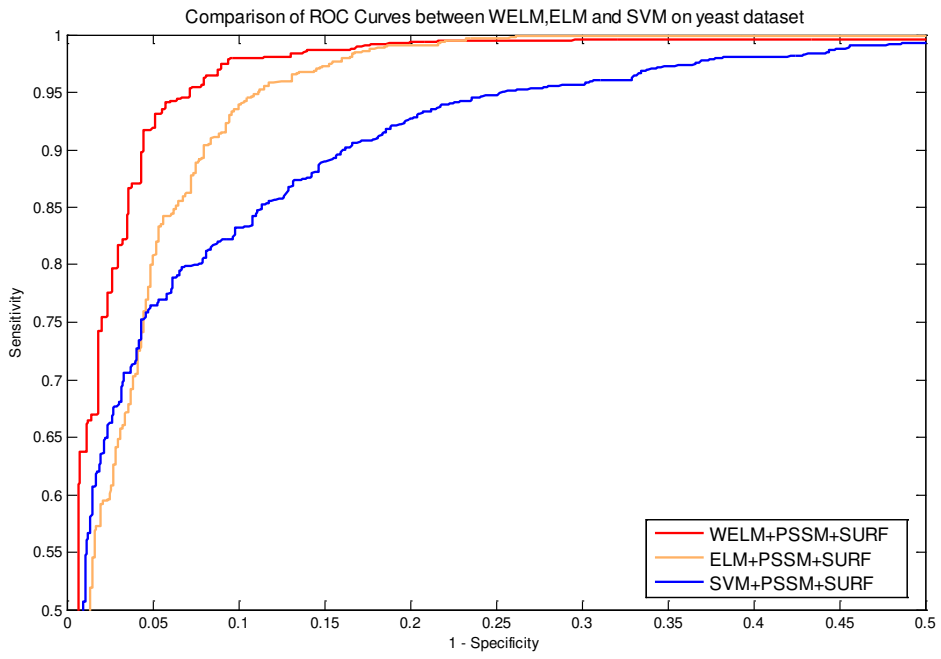

Figure 5 Comparison of ROC curves between WELM, ELM and SVM on *yeast* dataset.

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

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

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**

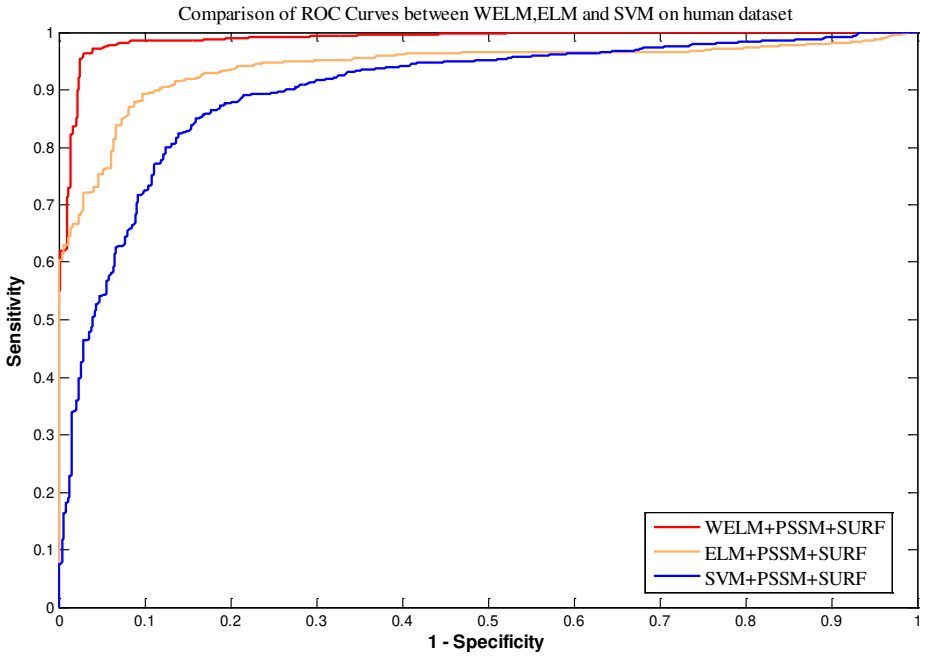

Figure 6 Comparison of ROC curves between WELM, ELM and SVM on *human* dataset.

3.3. Comparison with Other Methods

To further evaluate the prediction performance of WELM-SURF, we compare WELM-SURF method with the previous methods ,for example ,SLIPPER[36],CRS[28], SPAR[28] , DXECPPI , PPIevo [37] and LocFuse [38]. Table 7-8 displayed comparison results between different

prediction model on *yeast* and *human* dataset. It is easy to find from Table 7-8 that the prediction accuracy of WELM-SURF is significantly better than the other six prediction models on *yeas*t and *human* dataset. By comparing the results in Table 7-8, a similar conclusion that can be reached the proposed WELM-SURF method has very good predictive ability and can be used to high-quality predict SIPs. These comparison results further demonstrate the applicability of WELM-SURF forecasting SIP. This is mainly because the WELM is a robust and efficiently classifier and SURF can extract useful feature information of protein sequence. These comparison results further demonstrated that the WELM-SURF is suitable for identifying SIPs.

Table 7 Comparison results between WELM-SURF and other methods on *yeast* dataset

Model	Acc (%)	TNR (%)	TPR (%)	MCC
SLIPPER[36]	71.90	72.18	69.72	0.2842
PPIevo[37]	66.28	87.46	60.14	0.1801
LocFuse[38]	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	**95.25**	**85.79**	**93.05**	**0.8644**

Table 8 Comparison results between WELM-SURF and other methods on *human* dataset

Model	Acc (%)	TNR (%)	TPR (%)	MCC
SLIPPER[36]	91.10	95.06	47.26	0.4197
PPIevo[37]	78.04	25.82	87.83	0.2082
LocFuse[38]	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	**98.79**	**98.24**	**95.15**	**0.9189**

4. Conclusion

In the paper, we put forward a new computational method called WELM-SURF for predicting SIPs, which combines the Weighted Extreme Learning Machine (WELM) with Speeded up robust features (SURF) to predict SIPs based on evolutionary information of protein sequence. The experimental results proved that the proposed WELM-SURF model is competent for predicting SIPs with high accuracy and robustness and its prediction ability is significantly better than that 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) The PSSM contains not only the position information of the protein sequence, but also the evolution information that reflects the conservative function of protein and a number of prior information. Therefore, it can provide a certain help in extracting evolutionary information of protein sequence and capture key SIP features. (2) SURF can improve computational speed compared to SIFT. The main advantage of SURF that it 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 can capture self-protein interaction information and extract high efficiency features from PSSM. (3) For the sake of SIPs datasets are very imbalanced

class samples and the WELM has the advantage of short training time and good generalization ability and can efficiently execute classification for imbalanced class samples by optimizing the loss function of weight matrix. The WELM classifier can better perceive the distribution information of imbalanced class 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. Therefore, we can come to the conclusion that the proposed WELM-SURF model is useful tools and can execute incredibly well for predicting SIPs, as well as other bioinformatics tasks.

Abbreviations:

SIPs: Self-interaction Proteins
PPIs: Protein-protein interactions
WELM: Weighted Extreme Learning Machine
SIFT :Scale Invariant Feature Transform
SURF: Speeded Up Robust Features
PSSM: Position Specific Scoring Matrix
SVM: Support Vector Machine
ELM: Extreme Learning Machine
PSI-BLAST: Position-Specific Iterated BLAST
Acc: Accuracy
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： 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)'.The role the funder is Ji-Yong An who is corresponding author and first author.
Author Contributions: AJY and ZY conceived the algorithm, carried out analyses, prepared the data sets, carried out experiments, and wrote the manuscript; YZJ and ZYJ designed, performed and analyzed experiments and wrote the manuscript; all authors read 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* 2017, **9**(1):47.
11. 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-.
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* 2016, **17**(1):184.
13. 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.
14. 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.
15. Lei Z: **Sequence-Based Prediction of Protein-Protein Interactions Using Random Tree and Genetic Algorithm**. In: *International Conference on Intelligent Computing: 2012*.
16. Yang L, Xia JF, Gui J: **Prediction of protein-protein interactions from protein sequence using local descriptors**. *Protein & Peptide Letters* 2010, **17**(9):1085.
17. 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.

18. 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.
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. 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.
23. 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.
24. 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.
25. 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.
26. 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.
27. Consortium UP: **UniProt: a hub for protein information**. *Nucleic Acids Research* 2014, **43**(D1):D204-212.
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. Bay H, Tuytelaars T, Gool LV: **SURF: Speeded Up Robust Features**. 2006.
31. Lowe DG: **Object Recognition from Local Scale-Invariant Features**. In: *Computer Vision, 1999 The Proceedings of the Seventh IEEE International Conference on: 1999*.
32. Lowe DG: **Distinctive Image Features from Scale-Invariant Keypoints**. *International Journal of Computer Vision* 2004, **60**(2):91---110.
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. 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.
37. 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.
38. 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

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

Figure 1

the diagram of PSSM

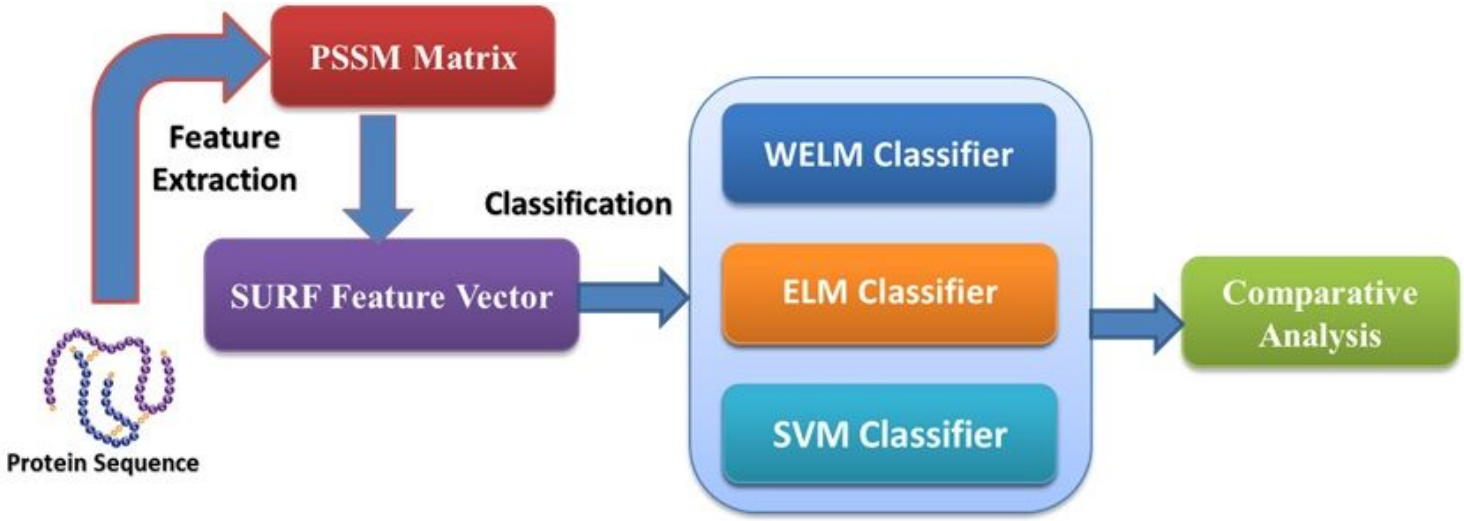

Figure 2

the flow diagram of our method

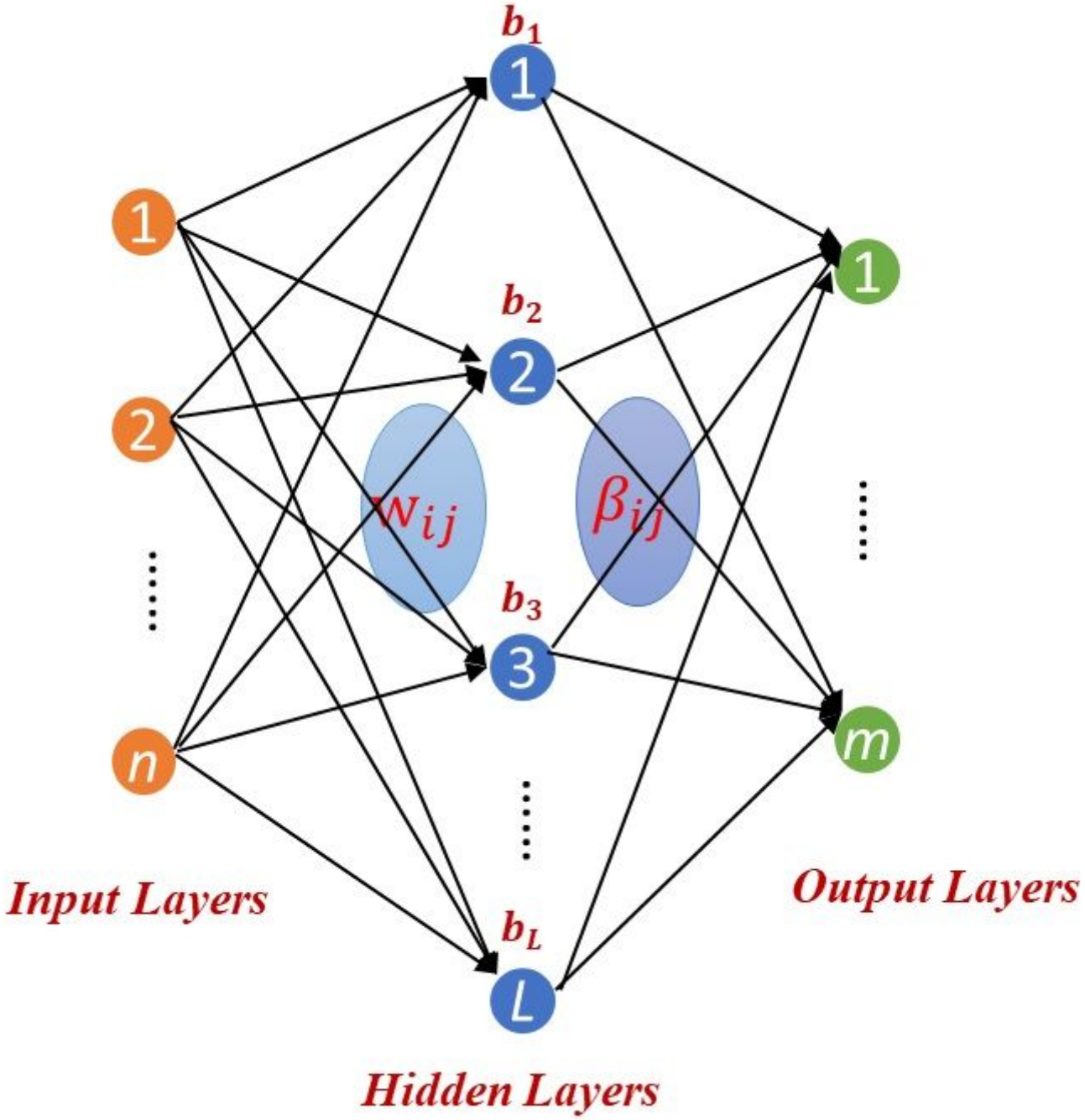

Figure 3

the network structure of ELM

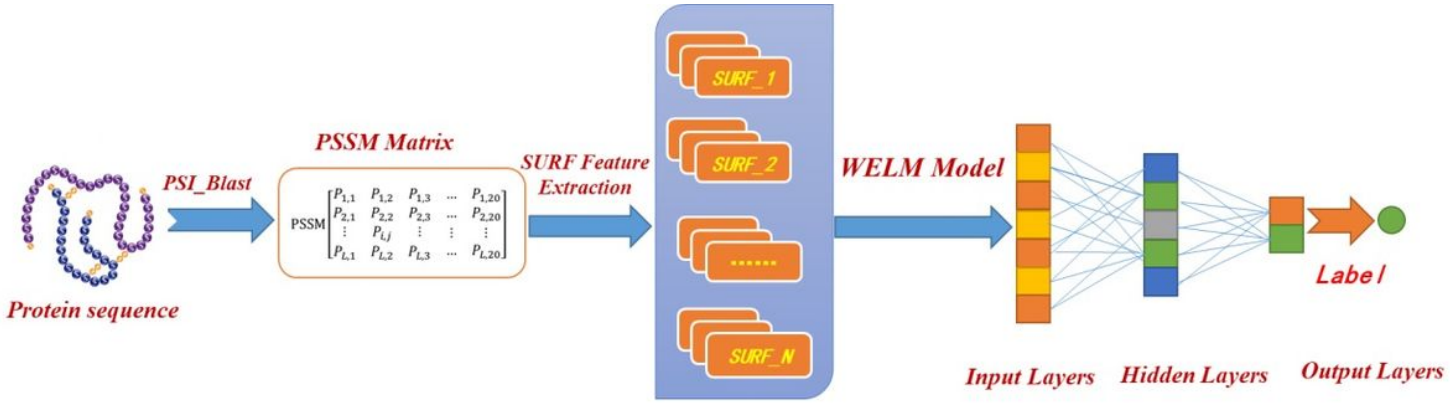

Figure 4

the prediction flow diagram of WELM-SURF

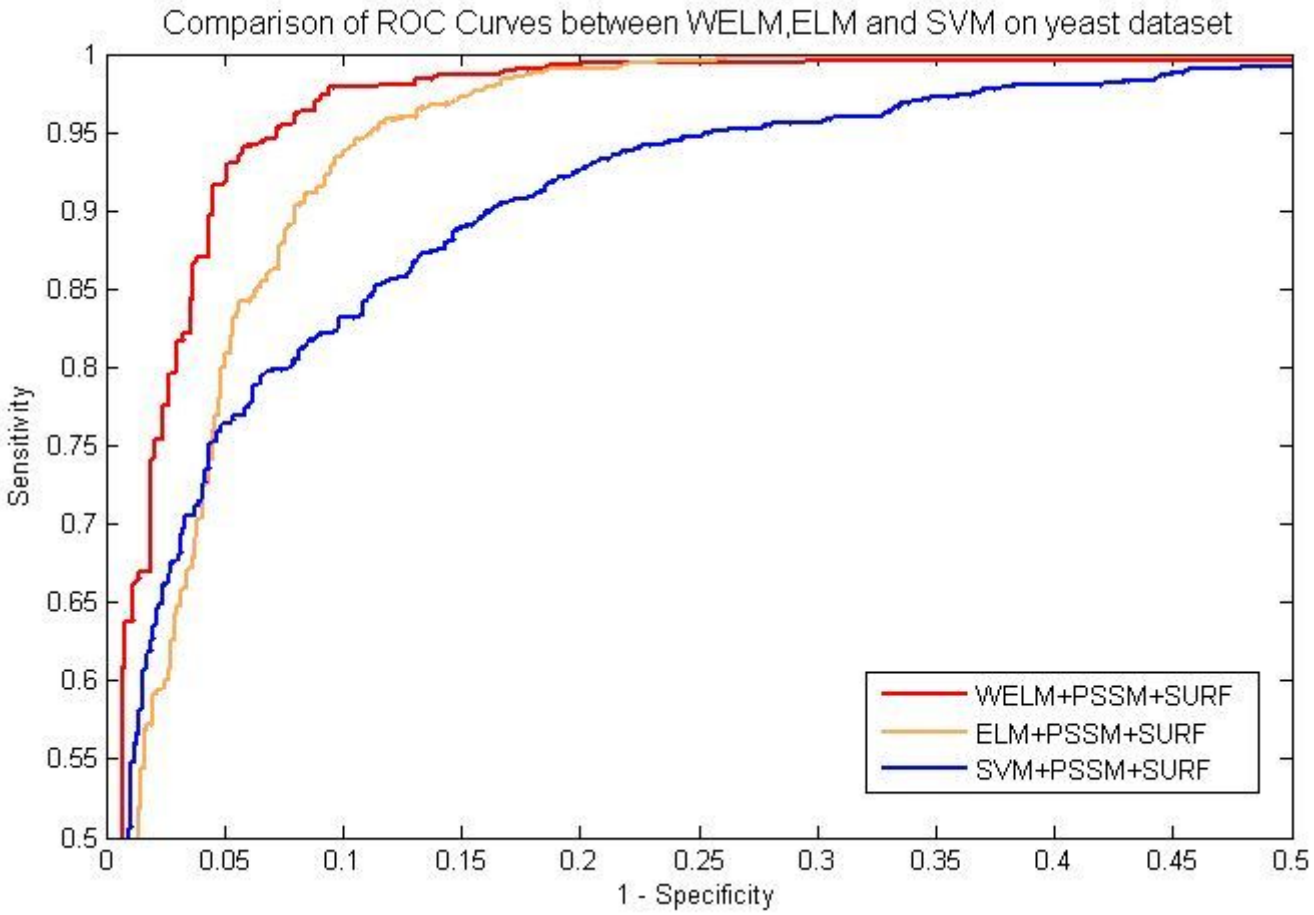

Figure 5

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

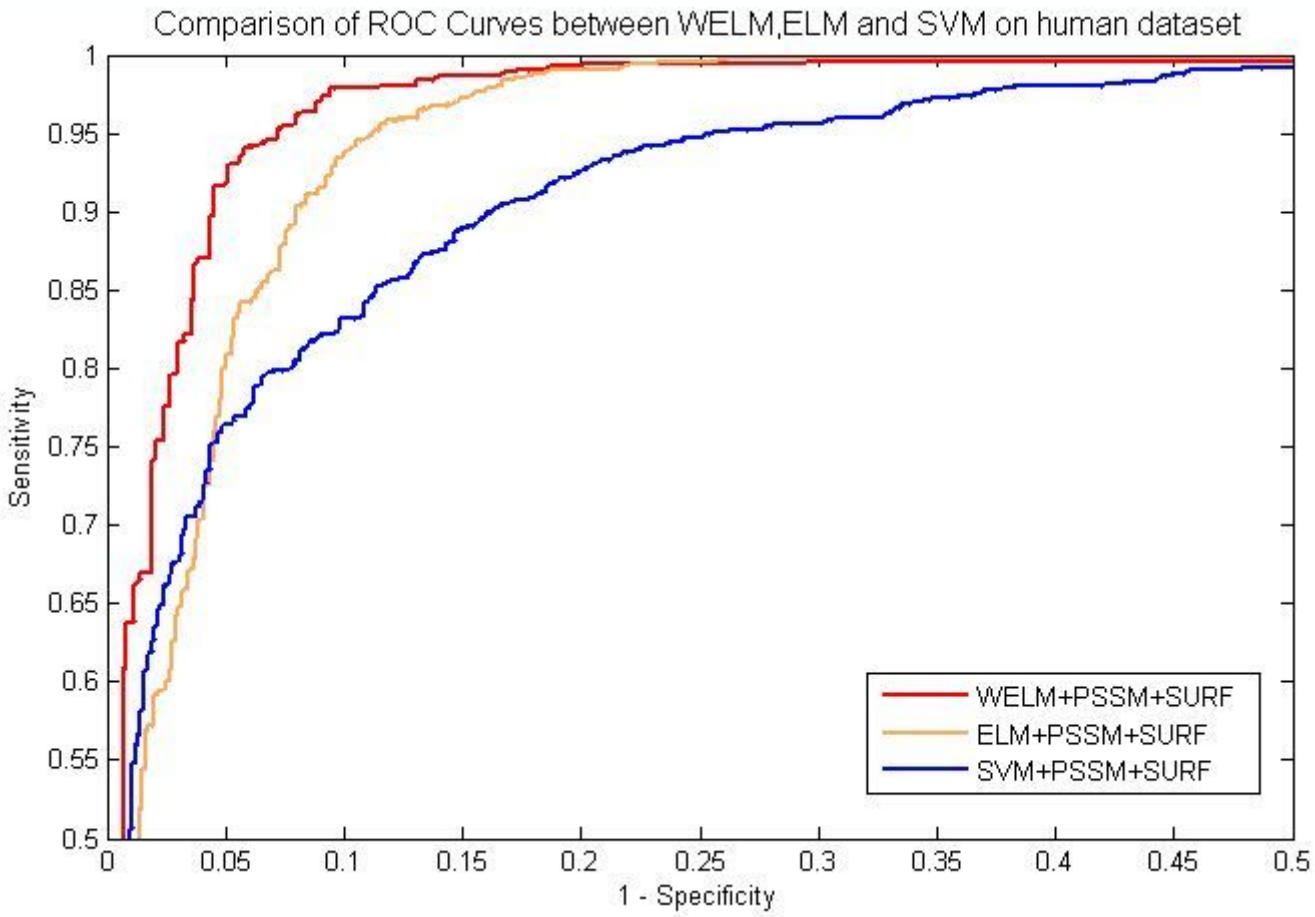

Figure 6

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