International Journal of
Molecular Sciences

MDPI

Article

Prediction of Self-Interacting Proteins from Protein Sequence Information Based on Random Projection Model and Fast Fourier Transform

Zhan-Heng Chen [1,2], Zhu-Hong You [1,2,*], Li-Ping Li [1], Yan-Bin Wang [1], Leon Wong [1,2] and Hai-Cheng Yi [1,2]

[1] The Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences, Urumqi 830011, China; chenzhanheng17@mails.ucas.ac.cn (Z.-H.C.); Lipingli@ms.xjb.ac.cn (L.-P.L.); wangyanbin15@mails.ucas.ac.cn (Y.-B.W.); huangliguang18@mails.ucas.ac.cn (L.W.); yihaicheng17@mails.ucas.ac.cn (H.-C.Y.)

[2] University of Chinese Academy of Sciences, Beijing 100049, China

* Correspondence: zhuhongyou@ms.xjb.ac.cn; Tel.: +86-181-6062-2862

Received: 5 December 2018; Accepted: 7 January 2019; Published: 21 February 2019

Abstract: It is significant for biological cells to predict self-interacting proteins (SIPs) in the field of bioinformatics. SIPs mean that two or more identical proteins can interact with each other by one gene expression. This plays a major role in the evolution of protein-protein interactions (PPIs) and cellular functions. Owing to the limitation of the experimental identification of self-interacting proteins, it is more and more significant to develop a useful biological tool for the prediction of SIPs from protein sequence information. Therefore, we propose a novel prediction model called RP-FFT that merges the Random Projection (RP) model and Fast Fourier Transform (FFT) for detecting SIPs. First, each protein sequence was transformed into a Position Specific Scoring Matrix (PSSM) using the Position Specific Iterated BLAST (PSI-BLAST). Second, the features of protein sequences were extracted by the FFT method on PSSM. Lastly, we evaluated the performance of RP-FFT and compared the RP classifier with the state-of-the-art support vector machine (SVM) classifier and other existing methods on the *human* and *yeast* datasets; after the five-fold cross-validation, the RP-FFT model can obtain high average accuracies of 96.28% and 91.87% on the *human* and *yeast* datasets, respectively. The experimental results demonstrated that our RP-FFT prediction model is reasonable and robust.

Keywords: self-interacting proteins; position-specific scoring matrix; fast Fourier transform; random projection

1. Introduction

Protein is an important component of all cells. It is an organic macromolecule and the basic material of life. It also is the main undertaker of activity. Without protein, there is no life. Most proteins often work together with a partner or other proteins. They can interact with two or more copies by themselves, which is termed self-interacting proteins (SIPs). However, for most researchers, whether proteins can interact with each other is a difficult thing to determine. SIPs play a key role in the development of protein interaction networks (PINs) [1,2]. The functions of many proteins, which could control the transport of ions and small molecules that pass through cell membranes, depends on their homo-oligomers [3]. Ispolatov et al. discovered that the average quantity of SIPs is more than twice that of other proteins in the PINs [4]. It is crucial for elucidating the functions of SIPs to comprehend whether a protein can self-interact; this also gives us an insight into the adjustment of protein function and can help us achieve a better comprehension of disease mechanisms [5]. Over the past few years,

many studies have shown that homo-oligomerization plays an important role in many biological processes, such as signal transduction, gene expression regulation, immune response, and enzyme activation [6–9]. Therefore, SIPs will be useful for improving steadiness and preventing against cellular stress and the denaturation of proteins via reducing the surface area [10].

So far, there are many ways to study bioinformatics [11–16] and genomics [17–22], and a number of previous methods for predicting PPIs have been put forward. For example, Pitre et al. [23] raised a new Protein-Protein Interaction Prediction Engine (PIPE), which could predict PPIs for any target pair of the *yeast S. cerevisiae* proteins from their primary structure and without the need for any additional information or predictions about the proteins. Xia et al. [24] put forward a sequence-based multi-classifier system that applied auto-correlation descriptor to encode a protein interaction pair and selected rotation forest as classifier to deduce PPIs. However, these methods are good for PPI detection [25] but have certain limitations in that they must take the correlation between protein pairs into account for Protein Self-interaction detection—for example, co-expression, co-localization, and co-evolution. Nevertheless, this information is useless for SIPs. Moreover, the datasets for PPI detection are balanced and those of SIPs are unbalanced. Besides, prediction of PPIs datasets has no PPIs between the same partners. For these reasons, the above computational models are not suitable for SIPs detection. Accordingly, it is becoming more and more crucial to exploit an effective calculation method to predict SIPs.

In our study, a random projection (RP) method for SIPs prediction from protein sequence information with Fast Fourier Transform (FFT) was proposed. Furthermore, the main idea of our proposed method includes four aspects: (1) the protein sequence information could be described as a Position-Specific Scoring Matrix (PSSM); (2) using the fast Fourier transform (FFT) method to extract eigenvectors from protein sequences on a PSSM; (3) using the Principal Component Analysis (PCA) approach to convert the high-dimensional data into useful information after FFT and the noise is removed, so the pattern in the data is found; (4) the RP algorithm is employed to build a training set where the classifier will be trained. Take it in detail as follows: first, the PSSM from each protein sequence is likely to result in a eigenvector whose dimension is 400 by applying the FFT method for extracting important information; then, reduce the dimension of the FFT vector to 300 for improving the performance of prediction by employing the PCA dimensionality reduction method; eventually, perform classification on *yeast* and *human* datasets by applying the RP classifier. The results demonstrate that this method outperforms the SVM-based approach and six other existing technologies. This indicates that the proposed model is suitable and performs well for predicting SIPs.

2. Results and Discussion

2.1. Performance Evaluation

In this study, to estimate the stability and availability of our prediction model, we used five measurements that were commonly used in binary classification tasks, including accuracy (Acc.), sensitivity (Sen.), specificity (Spe.), Matthews correlation coefficient (MCC) [26–32], and Balanced Accuracy (B_Acc.) [33], respectively. They could be defined as follows:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Sen = \frac{TP}{TP + FN} \tag{2}$$

$$Spe = \frac{TN}{TN + FP} \tag{3}$$

$$MCC = \frac{(TP \cdot TN) - (FP \cdot FN)}{\sqrt{(TP + FN)(TN + FP)(TP + FP)(TN + FN)}} \tag{4}$$

$$\text{B\_Acc} = \frac{Sen + Spe}{2} = \frac{2TP{\cdot}TN + TP{\cdot}FP + TN{\cdot}FN}{2(TP + FN)(TN + FP)}, \quad (5)$$

where TP represents the count of true positives, that is to say the number of real interacting pairs predicted correctly. FP is the quantity of false positives, defined as the volume of real non-interacting pairs mis-predicted. TN stands for the count of true negatives, which is the quantity of real non-interacting pairs correctly predicted. FN means the quantity of false negatives; in other words, it represents the true sample error predicted to be false samples. On the basis of these parameters, a Receiver Operating Curve (ROC) was plotted to assess the performance of the random projection approach. Then we can compute the area under the curve (AUC) to estimate the quality of the classifier.

2.2. Performance of the Proposed Method

In order to evaluate the performance of the presented model and avoid the overfitting problem, we applied the RP-FFT model to the *human* dataset. In statistical prediction, three cross-validation (CV) methods, such as an independent dataset test, a sub-sampling (or k-fold CV) test, and a leave-one-out CV (LOOCV) test, are frequently used to calculate the expected success rate of a developed predictor [34–38]. Among the three methods, however, the LOOCV test is deemed the least arbitrary and most objective, as demonstrated by Equations (28)–(32) of [39], and hence it has been widely recognized and increasingly adopted by investigators to examine the quality of various predictors [38,40,41]. However, it seems time- and resource-consuming. Thus, we used 5-fold CV to examine the proposed models. In 5-fold CV, the benchmarking dataset was randomly partitioned into 10 subsets. One subset is used as a test set and the remaining nine subsets are used as the training sets. This procedure is repeated five times, where each subset is used once as a test set. The performance of the five corresponding results is averaged to give the performance of the classifier. To assess the feasibility and stability of our prediction method, we also estimated the prediction performance of RP-FFT model on the *yeast* dataset.

To ensure the fairness of the experiment, we optimized a number of parameters for the RP-FFT prediction model. In this paper, we set up the same parameters for *human* and *yeast* datasets. Thus, we classify the training and test sets for $B1 = 10$ independent projections, each one carefully chosen from a block of size $B2 = 30$, and then chose the K-Nearest Neighbor (KNN) as the base classifier and the leave-one-out test error estimate, where $k = seq\ (1, 40, \text{by} = 3)$.

Our model can not only deal with balanced data, but can also solve the imbalanced data problem to some extent. At first, we employed the undersampling technique, as mentioned in [18], to solve the imbalanced dataset problem. The *human* dataset included 1441 SIPs as positives and 1441 non-SIPs as negatives. Using the same strategy, the *yeast* dataset contained 710 positive samples and 710 negative samples. The experimental results can be seen in Tables 1 and 2.

In addition, the initial imbalanced data collected from DIP, BioGRID, IntAct, InnateDB, and MatrixDB also used to compare our proposed method with previous work. If we use the undersampling technique to reconstruct the dataset, the size of the initial imbalanced data will be substantially reduced. As shown in Tables 2 and 3, we performed our proposed model on the initial imbalanced data in the experiment.

Table 1. The results of the RP-FFT method with 5-fold cross-validation on the *human* dataset.

Testing Set	Acc. (%)	Sen. (%)	Spe. (%)	MCC (%)
1	94.44	88.28	100.00	89.36
2	92.53	85.37	100.00	86.07
3	92.19	85.48	100.00	85.51
4	93.75	86.76	100.00	88.08
5	94.81	89.73	100.00	90.12
Average	93.54 ± 1.15	87.12 ± 1.87	100.00 ± 0.00	87.83 ± 2.01

Table 2. The results of the RP-FFT method with 5-fold cross-validation on the *yeast* dataset.

Testing Set	Acc. (%)	Sen. (%)	Spe. (%)	MCC (%)
1	80.99	97.12	65.52	65.71
2	83.45	92.14	75.00	68.03
3	82.04	97.89	66.20	67.57
4	84.86	95.14	74.29	71.13
5	83.45	92.41	74.10	67.83
Average	82.96 ± 1.48	94.94 ± 2.63	71.02 ± 4.73	68.05 ± 1.95

Table 3. The results of the RP-FFT method with 5-fold cross-validation on the *human* dataset.

Testing Set	Acc. (%)	Sen. (%)	Spe. (%)	MCC (%)	B_Acc. (%)
1	96.23	79.51	97.74	75.72	88.63
2	96.20	80.34	97.65	75.89	89.00
3	96.58	82.49	97.89	78.61	90.19
4	96.40	79.78	97.79	75.40	88.79
5	96.00	85.28	97.01	76.68	91.15
Average	96.28 ± 0.22	81.48 ± 2.43	97.62 ± 0.35	76.46 ± 1.29	89.55 ± 1.08

The experimental results of the RP-FFT prediction model on the *human* and *yeast* datasets are listed in Tables 3 and 4. Table 3 lists the data obtained that the model put forward obtained for average Accuracy (Acc.), Sensitivity (Sen.), Specificity (Spe.), Matthews correlation coefficient (MCC), and Balance accuracy (B_Acc.): 96.28%, 81.48%, 97.62%, 76.46%, and 89.55% for the *human* dataset and the standard deviations of them 0.22%, 2.43%, 0.35% 1.29%, and 1.08%, respectively. In the same way, we also got good results in Table 4 for average Acc., Sen., Spe., MCC and B_Acc.: 91.87%, 48.81%, 97.42%, 54.62% and 73.12%, and the standard deviations of them are 0.82%, 4.50%, 0.45%, 4.25%, and 2.30% for the *yeast* dataset, respectively.

Table 4. The results of the RP-FFT method with 5-fold cross-validation on the *yeast* dataset.

Testing Set	Acc. (%)	Sen. (%)	Spe. (%)	MCC (%)	B_Acc. (%)
1	91.32	50.00	96.73	53.09	73.37
2	91.72	47.33	97.81	55.35	72.57
3	92.20	49.63	97.39	54.80	73.51
4	91.00	42.36	97.36	49.06	69.86
5	93.09	54.74	97.83	60.82	76.29
Average	91.87 ± 0.82	48.81 ± 4.50	97.42 ± 0.45	54.62 ± 4.25	73.12 ± 2.30

From the above data, it is obvious that the proposed method could achieve good outcomes for SIPs predictions due to the suitable feature extraction and classifier. It can be summarized that the main improvement of our characteristic extraction technique contains the following factors: (1) The PSSM gives the score for finding a special matching amino acid in a target protein sequence. It is a good tool that can not only represent the protein sequence information but also saves enough prior information. Therefore, a PSSM contains all the major information of one protein sequence for detecting SIPs. (2) We extracted the features from the protein sequence by using the Fast Fourier Transform (FFT) method, which can further increase the performance of the RP-FFT model. (3) In case of ensuring the integrity information of FFT feature vector, we used Principal Component Analysis (PCA) to decrease the dimension of data and influence of noise, and thus the pattern in the data is found. Experimental results revealed that the eigenvector extracted from applying FFT on PSSM is quite suitable for SIP detection.

2.3. Comparison with Other Feature Extraction Methods

In this section, in order to illustrate the use of the FFT feature extraction method, we compared the FFT method with SVD (Singular Value Decomposition), DCT (Discrete Cosine Transform), and COV (Covariance) [42,43] on the Random Projection classifier. The results of RP classifier based on different feature extraction methods with 5-fold cross-validation on the *yeast* dataset are shown in Table 5. On the whole, it can be seen that the FFT feature extraction method works better than other methods for the *yeast* dataset.

Table 5. The results of RP classifier based on different feature extraction methods on the *yeast* dataset.

Feature Extraction Methods	Acc. (%)	Sen. (%)	Spe. (%)	MCC (%)	B_Acc. (%)
SVD	88.73 ± 0.75	10.25 ± 2.93	98.86 ± 0.43	19.76 ± 2.96	54.55 ± 1.31
DCT	90.35 ± 0.84	20.38 ± 2.62	99.36 ± 0.32	37.57 ± 1.74	59.87 ± 1.18
COV	91.93 ± 0.81	42.43 ± 4.82	98.31 ± 0.25	53.10 ± 4.91	70.37 ± 2.49
FFT	91.87 ± 0.82	48.81 ± 4.50	97.42 ± 0.45	54.62 ± 4.25	73.12 ± 2.30

2.4. Comparison with the SVM-Based Method

Though the RP-FFT model achieved better performance for predicting SIPs, we still need to further assess its use with our presented method. The veracity and stability of prediction of the RP classifier were compared with the state-of-the-art SVM method via the same characteristic extraction approach based on the *yeast* and *human* datasets, respectively. We applied the LIBSVM packet tool [44] to run the classification. Before the experiment, there are several parameters of SVM classifier should be optimized. In this paper, we chose a radial basis function (RBF) as the kernel function, and then used grid search to optimize the parameters of RBF, whose parameters were set to $c = 0.03$ and $g = 1200$.

As shown in Tables 6 and 7, we employed 5-fold cross-validation to train and compare the models of RP and SVM on the *yeast* and *human* datasets, respectively. The average Acc., the average Sen., the average Spe., the average MCC and B_Acc. of SVM classifier are 93.68%, 23.80%, 100.00%, 47.13%, and 61.90% on the *human* dataset in Table 6, respectively. Nevertheless, the RP classifier obtained 96.28% average Acc., 81.48% average Sen., 97.62% average Spe., 76.46% average MCC, and 89.55% average B_Acc. On the *human* dataset. Similarity, the average Accuracy, the average Sen., the average Spe., the average MCC and B_Acc. of SVM classifier are 90.63%, 17.79%, 100.00%, 39.95%, and 58.90% on the *yeast* dataset in Table 7. Nevertheless, the RP classifier received 91.87% average Acc., 48.81% average Sen., 97.42% average Spe., 54.62% average MCC and 73.12% average B_Acc. On the *human* dataset. In a word, it is obvious that the overall prediction result of RP classifier is much better than that of the SVM method.

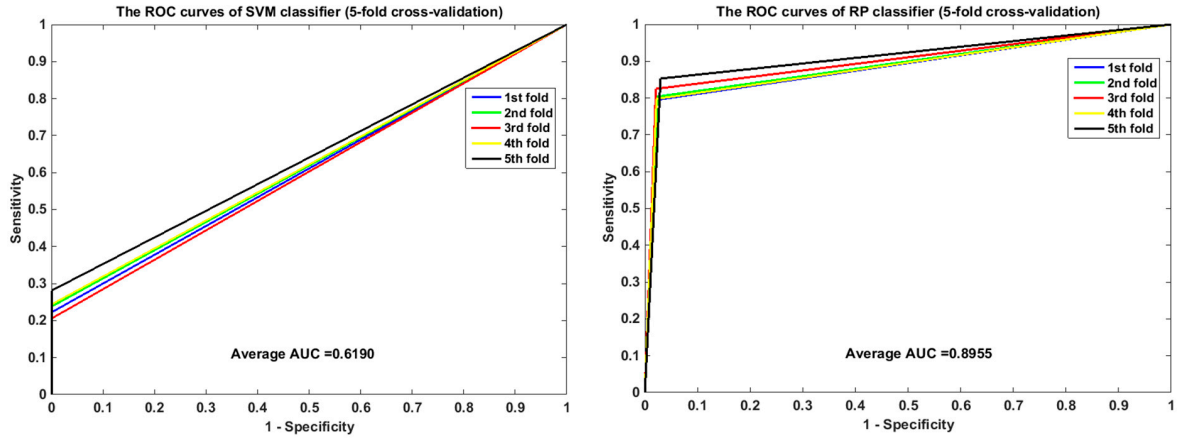
Table 6. Performance comparison of RP and SVM on the *human* dataset.

Model	Testing Set	Acc. (%)	Sen. (%)	Spe. (%)	MCC (%)	B_Acc. (%)
RP + FFT	1	96.23	79.51	97.74	75.72	88.63
	2	96.20	80.34	97.65	75.89	89.00
	3	96.58	82.49	97.89	78.61	90.19
	4	96.40	79.78	97.79	75.40	88.79
	5	96.00	85.28	97.01	76.68	91.15
	Average	96.28 ± 0.22	81.48 ± 2.43	97.62 ± 0.35	76.46 ± 1.29	89.55 ± 1.08
SVM + FFT	1	93.55	22.22	100.00	45.57	61.11
	2	93.64	23.79	100.00	47.17	61.90
	3	93.21	20.54	100.00	43.73	60.27
	4	94.19	24.34	100.00	47.86	62.17
	5	93.82	28.09	100.00	51.30	64.05
	Average	93.68 ± 0.36	23.80 ± 2.82	100.00 ± 0.00	47.13 ± 2.82	61.90 ± 1.41

Table 7. Performance comparison of RP and SVM on the *yeast* dataset.

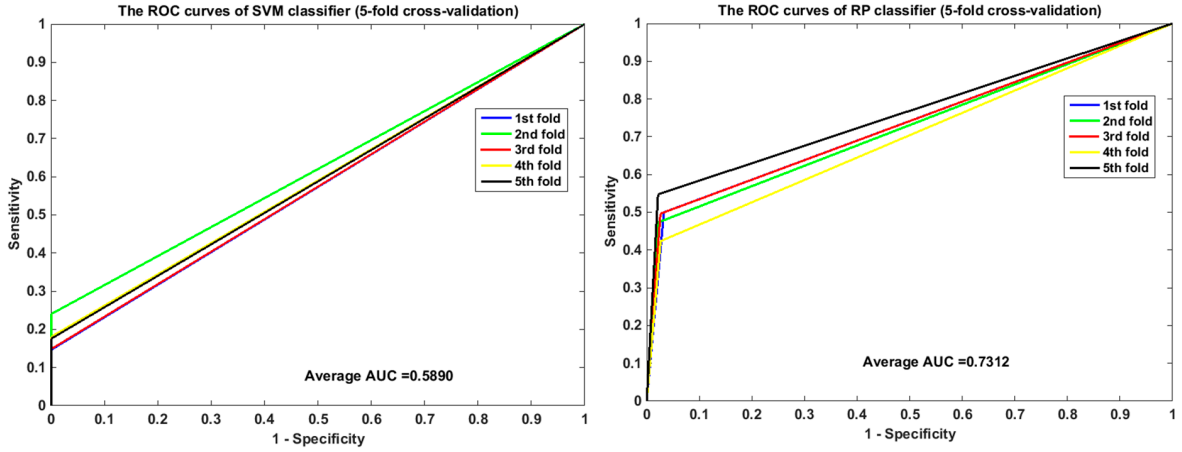
Model	Testing Set	Acc. (%)	Sen. (%)	Spe. (%)	MCC (%)	B_Acc. (%)
RP+FFT	1	91.32	50.00	96.73	53.09	73.37
	2	91.72	47.33	97.81	55.35	72.57
	3	92.20	49.63	97.39	54.80	73.51
	4	91.00	42.36	97.36	49.06	69.86
	5	93.09	54.74	97.83	60.82	76.29
	Average	91.87 ± 0.82	48.81 ± 4.50	97.42 ± 0.45	54.62 ± 4.25	73.12 ± 2.30
SVM+FFT	1	90.11	14.58	100.00	36.22	57.29
	2	90.84	24.00	100.00	46.62	62.00
	3	90.76	14.81	100.00	36.64	57.41
	4	90.51	18.06	100.00	40.38	59.03
	5	90.92	17.52	100.00	39.87	58.76
	Average	90.63 ± 0.33	17.79 ± 3.80	100.00 ± 0.00	39.95 ± 4.17	58.90 ± 1.90

Meanwhile, the ROC curves between RP and SVM on the *human* and *yeast* datasets are displayed in Figures 1 and 2. From Figure 1, it is clear that the average area under the curve (AUC) of SVM classifier is 0.6190 and that of the RP classifier is 0.8955. From Figure 2, we can see that the average AUC of SVM classifier is 0.5890 and that of the RP classifier is 0.7312. It is obvious that the average AUC of RP method is also larger than the AUC of the SVM method. So Random Projection is an accurate and robust method for SIP detection.



(**a**) ROC of SVM method (**b**) ROC of RP classifier

Figure 1. Comparison of ROC curves between RP and SVM on *human* (5-fold cross validation). (**a**) is the ROC curve of SVM method on *human* dataset by 5-fold cross validation. (**b**) is the ROC curve of RP classifier on *human* dataset by 5-fold cross validation.



(**a**) ROC of SVM method (**b**) ROC of RP classifier

Figure 2. Comparison of ROC curves between RP and SVM on *yeast* (5-fold cross validation). (**a**) is the ROC curve of SVM method on *yeast* dataset by 5-fold cross validation. (**b**) is the ROC curve of RP classifier on *yeast* dataset by 5-fold cross validation.

2.5. Comparison with Other Existing Methods

In our study, we compared the presented model, termed RP-FFT, with other existing models on the *yeast* and *human* datasets to further prove that it can achieve good results. These comparison results of RP-FFT models and other models on the *yeast* and *human* datasets are shown in Tables 8 and 9. From Table 8, it is obvious that the RP-FFT model obtained a higher average accuracy than other existing models on *yeast* dataset. It is also clear that the other six methods got lower specificity and sensitivity than our proposed model for the same dataset. Accordingly, as is apparent from Table 9, the overall outcomes of our prediction model are also significantly better than the other six models on the *human* dataset. To sum up, the experimental results of the proposed model called RP-FFT prove its accuracy for predicting SIPs compared with the six approaches. This explains why our prediction model is superior to the other six methods, because it employs a good method of feature extraction and a suitable classifier. It can be further illustrated that our RP-FFT model is suitable for predicting SIPs.

Table 8. Comparison of RP-FFT with the other existing models on the *yeast* dataset.

Model	Acc. (%)	Spe. (%)	Sen. (%)	MCC (%)	B_Acc. (%)
SLIPPER [2]	71.90	72.18	69.72	28.42	70.95
DXECPPI [45]	87.46	94.93	29.44	28.25	62.19
PPIevo [46]	66.28	87.46	60.14	18.01	73.80
LocFuse [47]	66.66	68.10	55.49	15.77	61.80
CRS [48]	72.69	74.37	59.58	23.68	66.98
SPAR [48]	76.96	80.02	53.24	24.84	66.63
Proposed method	91.87	97.42	48.81	54.62	73.12

Table 9. Comparison of RP-FFT with the other existing models on the *human* dataset.

Model	Acc. (%)	Spe. (%)	Sen. (%)	MCC (%)	B_Acc. (%)
SLIPPER [2]	91.10	95.06	47.26	41.97	71.16
DXECPPI [45]	30.90	25.83	87.08	8.25	56.46
PPIevo [46]	78.04	25.82	87.83	20.82	56.83
LocFuse [47]	80.66	80.50	50.83	20.26	65.67
CRS [48]	91.54	96.72	34.17	36.33	65.45
SPAR [48]	92.09	97.40	33.33	38.36	65.37
Proposed method	96.28	97.62	81.48	76.46	89.55

3. Materials and Methodology

3.1. Datasets

The datasets derived from the UniProt database [49] include 20,199 curated *human* protein sequences. The PPIs data could be collected from a variety of sources, including DIP [50], BioGRID [51], IntAct [52], InnateDB [53], and MatrixDB [54]. In this experiment, we mainly built the PPIs dataset, which obtains two identical interacting proteins and whose style of interaction was described as 'direct interaction' in correlative databases. On this foundation, 2994 *human* SIPs could be obtained.

We built the datasets to estimate the performance of our prediction method, which has three steps [48]: (1) protein sequences with a length less than 50 or more than 5000 residues from the *human* proteome were removed; (2) to build the *human* positive dataset, we picked out the SIPs data with high quality, which should meet one of the following requirements: (a) the self-interactions were discovered by at least one small-scale experiment or two types of large-scale experiments; (b) we annotated the protein as a homo-oligomer (comprising homodimer and homotrimer) in UniProt; (c) it has been reported by at least two publications for self-interactions; (3) for the *human* negative dataset, we eliminated SIPs from all the *human* proteome (containing proteins labeled as 'direct interaction' and much wider 'physical association') and the prediction of SIPs in the UniProt database. Eventually, the *human* dataset contained 1441 SIPs as a positive dataset and 15,938 non-SIPs as a negative dataset [48].

In addition, the *yeast* dataset was also built to further illustrate the cross-species performance of the RP-FFT model, which included 710 SIPs samples and 5511 non-SIPs samples [48] via the same strategy mentioned above.

3.2. Position-Specific Scoring Matrix

We discovered distantly correlative proteins by applying the Position-Specific Scoring Matrix (PSSM) [55–57], which is a helpful tool. Therefore, a PSSM can be transformed from each protein sequence information by applying the Position-Specific Iterated BLAST (PSI-BLAST) [58]. Then, each protein sequence could be transformed into an $N \times 20$ PSSM matrix as follows:

$$M = \left\{ M_{\alpha\beta},\ \ \alpha = 1, \cdots, N,\ \ \beta = 1, \cdots, 20 \right\}, \tag{6}$$

where N indicates the size of a protein sequence, and each protein gene was constructed by 20 types of amino acids. For the query protein sequence, a PSSM could arrange the value $M_{\alpha\beta}$ that represents the β-th amino acid at the position of α. Thus, $M_{\alpha\beta}$ could be described as:

$$M_{\alpha\beta} = \sum_{k=1}^{20} p(\alpha, k) \times q(\beta, k), \tag{7}$$

where $p(\alpha, k)$ means the occurrence frequency score of the k-th amino acid in the position of α with the probe, and $q(\beta, k)$ represents the value of Dayhoff's mutation matrix between the β-th and k-th amino acids. Accordingly, a high value is a strongly conservative position; otherwise, it means a weakly conservative position.

In conclusion, PSSM could be a helpful tool for predicting self-interacting proteins. Each PSSM from the protein sequence was generated by employing PSI-BLAST for SIPs detection. For the sake of getting a high degree and a wide range of homologous information, we chose three iterations and assigned the e-value of PSI-BLAST to be 0.001 in this process. Consequently, the PSSM of each protein sequence could be expressed as a matrix consisting of $M \times 20$ elements, where row M of the matrix means the quantity of residues of each protein, and column 20 of the PSSM indicates the 20 different kinds of amino acids.

3.3. Fast Fourier Transform

Fast Fourier Transform (FFT) [59] was first applied in digital signal processing in a number of diverse areas. Afterwards it was used for image processing for a given curve C whose shape was a closed scope. At a certain time t, there is a data sequence $F(t)$, $0 \leq t < T$. Since $F(t)$ is a periodic function, $F(t) = F(t + nT)$. In this study, we used it to extract the eigen values. Hence, we expand $F(t)$ into a Fourier series as much as possible; it can be described as follows:

$$F(t) = \sum_{-\infty}^{\infty} \omega_n e^{(2\alpha\pi nt/T)}, \tag{8}$$

where ω_n is the Fourier coefficients of $F(t)$.

$$\omega_n = \frac{1}{T}\int_0^T F(t)e^{(-2\alpha\pi nt/T)}dt, \qquad n \in Z \tag{9}$$

The discrete Fourier transform is given by

$$\omega_n = \frac{1}{N}\sum_{t=0}^{N-1} F(t)e^{(-2\alpha\pi nt/N)}, \quad n = 0, 1, \cdots, N-1, \tag{10}$$

where $\alpha = \sqrt{-1}$, $N = 2^n$, $n = 1, 2, \cdots, n_{\max}$. $F(t)$ is commonly named the shape signature, which represents the shape boundary of any one-dimensional function. Fourier transform could only capture the architectural characteristics of a shape, which is important to stem FFT from a perceptually meaningful shape signature. FFT stemmed from the centroid distance function is superior to FFT stemmed from other shape signatures. From the centroid (x_c, y_c) of the shape, the centroid distance function $r(t)$ could be defined by the distance of the boundary points:

$$r(t) = \left([x(t) - x_c]^2 + [y(t) - y_c]^2\right)^{1/2}, \tag{11}$$

where $x_c = \frac{1}{N}\sum_{t=0}^{N-1} x(t)$, $y_c = \frac{1}{N}\sum_{t=0}^{N-1} y(t)$ and N is the quantity of boundary points.

It is a matter of great significance to extract informative characteristics based on machine learning approaches. In our study, for the sake of each protein sequence being composed of amounts of amino acids, the eigenvector cannot be directly obtained from a PSSM by PSI-BLAST, which will lead to diverse length of eigenvectors. For solving the question, we multiply the transpose of PSSM by PSSM to obtain a 20×20 matrix, and the feature extraction method of fast Fourier transform is employed to generate characteristic vectors from the PSSM profile. In the end, each protein sequence could be calculated to a 400-dimensional vector after FFT. In this study, eventually, each protein sequence from the *yeast* and *human* datasets was transformed into a 400-dimensional vector by employing the fast Fourier transform method.

In our study, for the sake of obtaining the main important data and advancing the prediction accuracy, we used the Principal Component Analysis (PCA) approach to reduce the size of the *yeast* and *human* databases from 400 to 300. Furthermore, reducing the dimensionality of the datasets could remove the complexity of the classifier and improve the generalization error.

3.4. Support Vector Machine

Support vector machine (SVM) was first proposed by Cortes and Vapnik et al. [60] in 1995. SVM inherently do binary classification. SVM is a statistical learning theory method, which is mainly used in the field of pattern recognition. The purpose of SVM is to find the hyperplane that maximizes

the distance margin between the two classes. Hence, we can transform it into a convex quadratic programming problem. This idea can be expressed formally as follows:

$$\begin{array}{cc} \min\limits_{w,b,\xi} & \frac{w^T w}{2} + C\sum\limits_{i=1}^{l}\xi_i \\ \text{subject to} & y_i(w^T \varnothing(x_i) + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0 \end{array} \tag{12}$$

where (x_i, y_i) is a training set of instance-label pairs, $i = 1, ..., l$. $x_i \in R^n$ are mapped into a higher dimensional space by the function Ø. $y \in \{1, -1\}^l$. Furthermore, the kernel function can be described as $K(x_i, x_j) \equiv \varnothing(x_i)^T \varnothing(x_j)$. It has four basic kernels that can be found in [61]:

(1) Linear: $K(x_i, x_j) = x_i^T x_j$.
(2) Polynomial: $K(x_i, x_j) = (\gamma x_i^T x_j + r)^d$, $\gamma > 0$.
(3) Radial basis function (RBF): $K(x_i, x_j) = \exp(-\gamma \| x_i - x_j \|^2)$, $\gamma > 0$.
(4) Sigmoid: $K(x_i, x_j) = \tan h(\gamma x_i^T x_j + r)$.

Here, γ, r, and d are kernel parameters. In our experiment, we chose RBF as the kernel function.

3.5. Random Projection Classifier

In mathematics and statistics, random projection (RP) is a technique for dimensionality reduction of some points that exist in Euclidean space. The meaning of the RP method is that projecting N points in N dimensional space can almost always onto a space of dimension *ClogN* with control on the ratio of distances and the error [62]. This method has been successfully applied for the reestablishment of frequency-sparse signals [63,64], facial recognition [65–67], protein mapping [68], and textual and visual information retrieval [69].

Next, we formally describe the random projection technique in detail. First, let

$$\Gamma = \{A_i\}_{i=1}^{N}, A_i \in R^n \tag{13}$$

be the primitive high-dimensional space dataset, where n is the quantity of high dimension and N is the count of the dataset. The goal of descending dimension is embedding the eigenvectors into a lower dimensional space R^q from a high-dimensional R^n where $q << n$. The output of data is represented as follows:

$$\widetilde{\Gamma} = \left\{\widetilde{A}_i\right\}_{i=1}^{N}, \widetilde{A}_i \in R^q, \tag{14}$$

where q approaches the inherent dimensionality of Γ. Thus, the vectors of Γ were regarded as embedding vectors.

If we want to reduce the dimension of Γ via the random projection method, a random vector set $\gamma = \{r_i\}_{i=1}^{k}$ must first be constructed, where $r_i \in R^q$. The random basis can be obtained by two common choices, as follows [62]:

(1) The vectors $\{r_i\}_{i=1}^{k}$ are normally distributed over the q dimensional unit sphere.
(2) The components of the vectors $\{r_i\}_{i=1}^{k}$ are selected Bernoulli $+1/-1$ distribution and the vectors are standardized so that $\| r_i \|_{l2} = 1$ for $i = 1, \ldots, n$.

The columns of $q \times n$ matrix R consist of the vectors in γ. The embedding result $\tilde{A}_i$ of A_i can be got by

$$\widetilde{A}_i = R \cdot A_i \tag{15}$$

In our proposed method, random projection is employed to build a training set where the classifier would be trained. We enrich the component of the integration method by using random projection.

Next, the dimension of the objective space was set to one part around the space where the training members reside. We built a size of $n \times N$ matrix G whose columns are made up the column eigenvectors in Γ. The training set Γ is given in Equation (7).

$$G = (A_1|A_2|\ldots|A_N) \tag{16}$$

Then, we construct k random matrices $\{R_i\}_{i=1}^{k}$ whose magnitude is $q \times n$, q and n are mentioned in the above paragraph, and k is the quantity of integration classifiers. Here, the columns of matrices are normalized so the l_2 norm is 1.

Then, using our method, we constructed training sets $\{T_i\}_{i=1}^{k}$ by projecting G onto $\{R_i\}_{i=1}^{k}$ which is the k random matrix. It can be represented as follows:

$$T_i = R_i \cdot G, i = 1, \ldots, k. \tag{17}$$

The training sets are imported into an inducer and the export results are a set of classifiers $\{\ell_i\}_{i=1}^{k}$. How do we classify a new dataset I through classifier ℓ_i? First, we embed I into the dimensionality reduction space R^q. Then, it can be owned via embedding u in the random matrix R_i as follows:

$$\tilde{I} = R_i \cdot I, \tag{18}$$

where $\tilde{I}$ is the inlaying of u, the classification of $\tilde{I}$ can be garnered from the classification of I by ℓ_i. In this ensemble method, the random projection classifier apply a data-driven voting threshold that is employed on the classification results of the whole classifier $\{\ell_i\}_{i=1}^{k}$ for the $\tilde{I}$ to decide produce the ultimate classification result of $\tilde{I}$.

In this experiment, the random projections were segmented into non-overlapping parts, where $B1$ = 10 and each one was carefully chosen from a certain part of size $B2$ = 30 that achieved the smallest estimate of the test error. We chose the k-Nearest Neighbor (KNN) as the base classifier and the leave-one-out test error estimate, where $k = seq$ (1, 40, by = 3). The prior probability of interaction pairs in the training sample dataset was taken as the voting parameter. Our classifier integrates the results of taking advantage of the base classifier on the chosen projection, with the data-driven voting threshold confirming the ultimate mission.

4. Conclusions

In our study, we developed a new prediction model based on protein sequence information to detect SIPs. This model was created by combining Position-Specific Scoring Matrix with Fast Fourier Transform and Random Projection classifier, which was termed RP-FFT. The main point of the experiment is that the datasets used by the classifier are unbalanced. The main improvements of the presented model are: (1) making use of a reasonable feature extraction method that could capture the main information of the data to improve the performance efficiency. (2) The RP classifier is strongly suitable for SIPs prediction. To summarize, the experimental results achieved by the presented method on the *yeast* and *human* datasets indicated that our prediction performance is obviously better than that of the SVM-based method and six other existing models. In the future, there will be more and more characteristic extraction techniques and machine learning or deep learning methods attempted for detecting SIPs.

Author Contributions: Conceptualization, Z.-H.C. and Z.-H.Y.; methodology, L.-P.L.; software, Z.-H.C.; validation, Z.-H.C., Y.-B.W. and H.-C.Y.; formal analysis, L.W.; investigation, Y.-B.W.; resources, L.-P.L.; data curation, Z.-H.Y.; writing—original draft preparation, Z.-H.C.; writing—review and editing, Y.-B.W.; visualization, L.W.; supervision, L.-P.L.; project administration, H.-C.Y.; funding acquisition, Z.-H.Y.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 61373086. The authors would like to thank all the guest editors and anonymous reviewers for their constructive advice.

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

References

1. Chen, Z.-H.; You, Z.-H.; Li, L.-P.; Wang, Y.-B.; Li, X. RP-FIRF: Prediction of Self-interacting Proteins Using Random Projection Classifier Combining with Finite Impulse Response Filter. In Proceedings of the International Conference on Intelligent Computing, Wuhan, China, 15–18 August 2018; pp. 232–234.
2. 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. *Mol. Cell. Proteom.* **2013**. [CrossRef] [PubMed]
3. Marianayagam, N.J.; Sunde, M.; Matthews, J.M. The power of two: Protein dimerization in biology. *Trends Biochem. Sci.* **2004**, *29*, 618–625. [CrossRef] [PubMed]
4. Ispolatov, I.; Yuryev, A.; Mazo, I.; Maslov, S. Binding properties and evolution of homodimers in protein–protein interaction networks. *Nucleic Acids Res.* **2005**, *33*, 3629–3635. [CrossRef] [PubMed]
5. Wang, Y.-B.; You, Z.-H.; Li, L.-P.; Huang, Y.-A.; Yi, H.-C. Detection of interactions between proteins by using legendre moments descriptor to extract discriminatory information embedded in pssm. *Molecules* **2017**, *22*, 1366. [CrossRef] [PubMed]
6. Woodcock, J.M.; Murphy, J.; Stomski, F.C.; Berndt, M.C.; Lopez, A.F. The dimeric versus monomeric status of 14-3-3ζ is controlled by phosphorylation of Ser58 at the dimer interface. *J. Biol. Chem.* **2003**, *278*, 36323–36327. [CrossRef]
7. Baisamy, L.; Jurisch, N.; Diviani, D. Leucine zipper-mediated homo-oligomerization regulates the Rho-GEF activity of AKAP-Lbc. *J. Biol. Chem.* **2005**, *280*, 15405–15412. [CrossRef] [PubMed]
8. Katsamba, P.; Carroll, K.; Ahlsen, G.; Bahna, F.; Vendome, J.; Posy, S.; Rajebhosale, M.; Price, S.; Jessell, T.; Ben-Shaul, A. Linking molecular affinity and cellular specificity in cadherin-mediated adhesion. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 11594–11599. [CrossRef] [PubMed]
9. 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 Sci.* **2009**, *18*, 2060–2066. [CrossRef]
10. Miller, S.; Lesk, A.M.; Janin, J.; Chothia, C. The accessible surface area and stability of oligomeric proteins. *Nature* **1987**, *328*, 834. [CrossRef]
11. Zeng, X.; Liao, Y.; Liu, Y.; Zou, Q. Prediction and validation of disease genes using HeteSim Scores. *IEEE/ACM Trans. Comput. Biol. Bioinform. (TCBB)* **2017**, *14*, 687–695. [CrossRef]
12. Zou, Q.; Wan, S.; Ju, Y.; Tang, J.; Zeng, X. Pretata: Predicting TATA binding proteins with novel features and dimensionality reduction strategy. *BMC Syst. Biol.* **2016**, *10*, 114. [CrossRef]
13. Nanni, L.; Lumini, A.; Brahnam, S. A set of descriptors for identifying the protein–drug interaction in cellular networking. *J. Theor. Biol.* **2014**, *359*, 120–128. [CrossRef]
14. Nanni, L.; Brahnam, S. Set of approaches based on 3D structure and Position Specific Scoring Matrix for predicting DNA-binding proteins. *Bioinformatics* **2018**. [CrossRef]
15. You, Z.-H.; Huang, Z.-A.; Zhu, Z.; Yan, G.-Y.; Li, Z.-W.; Wen, Z.; Chen, X. PBMDA: A novel and effective path-based computational model for miRNA-disease association prediction. *PLoS Comput. Biol.* **2017**, *13*, e1005455. [CrossRef]
16. You, Z.-H.; Lei, Y.-K.; Gui, J.; Huang, D.-S.; Zhou, X. Using manifold embedding for assessing and predicting protein interactions from high-throughput experimental data. *Bioinformatics* **2010**, *26*, 2744–2751. [CrossRef]
17. Zou, Q.; Li, J.; Song, L.; Zeng, X.; Wang, G. Similarity computation strategies in the microRNA-disease network: A survey. *Brief. Funct. Genom.* **2015**, *15*, 55–64. [CrossRef]
18. Manavalan, B.; Shin, T.H.; Kim, M.O.; Lee, G. PIP-EL: A new ensemble learning method for improved proinflammatory peptide predictions. *Front. Immunol.* **2018**, *9*, 1783. [CrossRef]
19. Wang, Y.-B.; You, Z.-H.; Li, X.; Jiang, T.-H.; Cheng, L.; Chen, Z.-H. Prediction of protein self-interactions using stacked long short-term memory from protein sequences information. *BMC Syst. Biol.* **2018**, *12*, 129. [CrossRef]
20. Yi, H.-C.; You, Z.-H.; Huang, D.-S.; Li, X.; Jiang, T.-H.; Li, L.-P. A Deep Learning Framework for Robust and Accurate Prediction of ncRNA-Protein Interactions Using Evolutionary Information. *Mol. Ther. Nucleic Acids* **2018**, *11*, 337–344. [CrossRef]
21. You, Z.-H.; Zhou, M.; Luo, X.; Li, S. Highly efficient framework for predicting interactions between proteins. *IEEE Trans. Cybern.* **2017**, *47*, 731–743. [CrossRef]

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. *J. Theor. Biol.* **2017**, *418*, 105–110. [CrossRef] [PubMed]
23. Pitre, S.; Dehne, F.; Chan, A.; Cheetham, J.; Duong, A.; Emili, A.; Gebbia, M.; Greenblatt, J.; Jessulat, M.; Krogan, N. PIPE: A protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs. *BMC Bioinform.* **2006**, *7*, 365. [CrossRef] [PubMed]
24. Xia, J.-F.; Han, K.; Huang, D.-S. Sequence-based prediction of protein-protein interactions by means of rotation forest and autocorrelation descriptor. *Protein Pept. Lett.* **2010**, *17*, 137–145. [CrossRef] [PubMed]
25. Wang, Y.-B.; You, Z.-H.; Li, X.; Jiang, T.-H.; Chen, X.; Zhou, X.; Wang, L. Predicting protein–protein interactions from protein sequences by a stacked sparse autoencoder deep neural network. *Mol. BioSyst.* **2017**, *13*, 1336–1344. [CrossRef] [PubMed]
26. Basith, S.; Manavalan, B.; Shin, T.H.; Lee, G. iGHBP: Computational identification of growth hormone binding proteins from sequences using extremely randomised tree. *Comput. Struct. Biotechnol. J.* **2018**, *16*, 412–420. [CrossRef] [PubMed]
27. Manavalan, B.; Subramaniyam, S.; Shin, T.H.; Kim, M.O.; Lee, G. Machine-learning-based prediction of cell-penetrating peptides and their uptake efficiency with improved accuracy. *J. Proteome Res.* **2018**, *17*, 2715–2726. [CrossRef] [PubMed]
28. Wei, L.; Hu, J.; Li, F.; Song, J.; Su, R.; Zou, Q. Comparative analysis and prediction of quorum-sensing peptides using feature representation learning and machine learning algorithms. *Brief. Bioinform.* **2018**. [CrossRef]
29. Manavalan, B.; Shin, T.H.; Kim, M.O.; Lee, G. AIPpred: Sequence-Based Prediction of Anti-inflammatory Peptides Using Random Forest. *Front. Pharmacol.* **2018**, *9*, 276. [CrossRef]
30. Wei, L.; Luan, S.; Nagai, L.A.E.; Su, R.; Zou, Q. Exploring sequence-based features for the improved prediction of DNA N4-methylcytosine sites in multiple species. *Bioinformatics* **2018**. [CrossRef]
31. Manavalan, B.; Govindaraj, R.G.; Shin, T.H.; Kim, M.O.; Lee, G. iBCE-EL: A new ensemble learning framework for improved linear B-cell epitope prediction. *Front. Immunol.* **2018**, *9*, 1695. [CrossRef]
32. Wei, L.; Chen, H.; Su, R. M6APred-EL: A sequence-based predictor for identifying N6-methyladenosine sites using ensemble learning. *Mol. Ther. Nucleic Acids* **2018**, *12*, 635–644. [CrossRef]
33. Gabere, M.N.; Noble, W.S. Empirical comparison of web-based antimicrobial peptide prediction tools. *Bioinformatics* **2017**, *33*, 1921–1929. [CrossRef]
34. Manavalan, B.; Shin, T.H.; Lee, G. PVP-SVM: Sequence-based prediction of phage virion proteins using a support vector machine. *Front. Microbiol.* **2018**, *9*, 476. [CrossRef]
35. Wei, L.; Zhou, C.; Chen, H.; Song, J.; Su, R. ACPred-FL: A sequence-based predictor based on effective feature representation to improve the prediction of anti-cancer peptides. *Bioinformatics* **2018**, *34*, 4007–4016. [CrossRef]
36. Manavalan, B.; Shin, T.H.; Lee, G. DHSpred: Support-vector-machine-based human DNase I hypersensitive sites prediction using the optimal features selected by random forest. *Oncotarget* **2018**, *9*, 1944. [CrossRef]
37. Wei, L.; Tang, J.; Zou, Q. SkipCPP-Pred: An improved and promising sequence-based predictor for predicting cell-penetrating peptides. *BMC Genom.* **2017**, *18*, 1. [CrossRef]
38. Manavalan, B.; Basith, S.; Shin, T.H.; Choi, S.; Kim, M.O.; Lee, G. MLACP: Machine-learning-based prediction of anticancer peptides. *Oncotarget* **2017**, *8*, 77121. [CrossRef]
39. Chou, K.-C. Some remarks on protein attribute prediction and pseudo amino acid composition. *J. Theor. Biol.* **2011**, *273*, 236–247. [CrossRef]
40. Dao, F.-Y.; Lv, H.; Wang, F.; Feng, C.-Q.; Ding, H.; Chen, W.; Lin, H. Identify origin of replication in Saccharomyces cerevisiae using two-step feature selection technique. *Bioinformatics* **2018**. [CrossRef]
41. Manavalan, B.; Lee, J. SVMQA: Support–vector-machine-based protein single-model quality assessment. *Bioinformatics* **2017**, *33*, 2496–2503. [CrossRef]
42. Nanni, L.; Lumini, A.; Brahnam, S. An empirical study of different approaches for protein classification. *Sci. World J.* **2014**, *2014*, 236717. [CrossRef] [PubMed]
43. Nanni, L.; Brahnam, S.; Lumini, A. Wavelet images and Chou's pseudo amino acid composition for protein classification. *Amino Acids* **2012**, *43*, 657–665. [CrossRef] [PubMed]
44. Chang, C.-C.; Lin, C.-J. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol. (TIST)* **2011**, *2*, 27. [CrossRef]

45. Du, X.; Cheng, J.; Zheng, T.; Duan, Z.; Qian, F. A novel feature extraction scheme with ensemble coding for protein–protein interaction prediction. *Int. J. Mol. Sci.* **2014**, *15*, 12731–12749. [CrossRef] [PubMed]
46. 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*, 237–242. [CrossRef] [PubMed]
47. Zahiri, J.; Mohammad-Noori, M.; Ebrahimpour, R.; Saadat, S.; Bozorgmehr, J.H.; Goldberg, T.; Masoudi-Nejad, A. LocFuse: Human protein–protein interaction prediction via classifier fusion using protein localization information. *Genomics* **2014**, *104*, 496–503. [CrossRef] [PubMed]
48. 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*, 1655–1665. [CrossRef]
49. Consortium, U. UniProt: A hub for protein information. *Nucleic Acids Res.* **2014**, *43*, D204–D212. [CrossRef]
50. Salwinski, L.; Miller, C.S.; Smith, A.J.; Pettit, F.K.; Bowie, J.U.; Eisenberg, D. The database of interacting proteins: 2004 update. *Nucleic Acids Res.* **2004**, *32*, D449–D451. [CrossRef]
51. Chatr-Aryamontri, A.; Oughtred, R.; Boucher, L.; Rust, J.; Chang, C.; Kolas, N.K.; O'Donnell, L.; Oster, S.; Theesfeld, C.; Sellam, A. The BioGRID interaction database: 2017 update. *Nucleic Acids Res.* **2017**, *45*, D369–D379. [CrossRef]
52. Orchard, S.; Ammari, M.; Aranda, B.; Breuza, L.; Briganti, L.; Broackes-Carter, F.; Campbell, N.H.; Chavali, G.; Chen, C.; Del-Toro, N. The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* **2013**, *42*, D358–D363. [CrossRef] [PubMed]
53. Breuer, K.; Foroushani, A.K.; Laird, M.R.; Chen, C.; Sribnaia, A.; Lo, R.; Winsor, G.L.; Hancock, R.E.; Brinkman, F.S.; Lynn, D.J. InnateDB: Systems biology of innate immunity and beyond—recent updates and continuing curation. *Nucleic Acids Res.* **2012**, *41*, D1228–D1233. [CrossRef] [PubMed]
54. Chautard, E.; Fatoux-Ardore, M.; Ballut, L.; Thierry-Mieg, N.; Ricard-Blum, S. MatrixDB, the extracellular matrix interaction database. *Nucleic Acids Res.* **2010**, *39*, D235–D240. [CrossRef] [PubMed]
55. Gribskov, M.; McLachlan, A.D.; Eisenberg, D. Profile analysis: Detection of distantly related proteins. *Proc. Natl. Acad. Sci. USA* **1987**, *84*, 4355–4358. [CrossRef] [PubMed]
56. Wang, Y.; You, Z.; Li, X.; Chen, X.; Jiang, T.; Zhang, J. PCVMZM: Using the Probabilistic Classification Vector Machines Model Combined with a Zernike Moments Descriptor to Predict Protein–Protein Interactions from Protein Sequences. *Int. J. Mol. Sci.* **2017**, *18*, 1029. [CrossRef] [PubMed]
57. Wang, Y.-B.; You, Z.-H.; Li, L.-P.; Huang, D.-S.; Zhou, F.-F.; Yang, S. Improving Prediction of Self-interacting Proteins Using Stacked Sparse Auto-Encoder with PSSM profiles. *Int. J. Biol. Sci.* **2018**, *14*, 983–991. [CrossRef] [PubMed]
58. Altschul, S.F.; Koonin, E.V. Iterated profile searches with PSI-BLAST—A tool for discovery in protein databases. *Trends Biochem. Sci.* **1998**, *23*, 444–447. [CrossRef]
59. Ahmed, N.; Rao, K.R. *Orthogonal Transforms for Digital Signal Processing*; Springer Science & Business Media: Berlin, Germany, 2012.
60. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
61. Hsu, C.-W.; Chang, C.-C.; Lin, C.-J. *A Practical Guide to Support Vector Classification*; National Taiwan University: Taipei, Taiwan, 2003.
62. Schclar, A.; Rokach, L. Random projection ensemble classifiers. In Proceedings of the International Conference on Enterprise Information Systems, Milan, Italy, 6–10 May 2009; pp. 309–316.
63. Candès, E.J.; Romberg, J.; Tao, T. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inf. Theory* **2006**, *52*, 489–509. [CrossRef]
64. Donoho, D.L. Compressed sensing. *IEEE Trans. Inf. Theory* **2006**, *52*, 1289–1306. [CrossRef]
65. Goel, N.; Bebis, G.; Nefian, A. Face recognition experiments with random projection. *Proc. SPIE* **2005**, *5779*, 426–438.
66. Lumini, A.; Nanni, L.; Brahnam, S. Ensemble of texture descriptors and classifiers for face recognition. *Appl. Comput. Inf.* **2017**, *13*, 79–91. [CrossRef]
67. Nanni, L.; Lumini, A.; Brahnam, S. Ensemble of texture descriptors for face recognition obtained by varying feature transforms and preprocessing approaches. *Appl. Soft Comput.* **2017**, *61*, 8–16. [CrossRef]

68. Linial, M.; Linial, N.; Tishby, N.; Yona, G. Global self-organization of all known protein sequences reveals inherent biological signatures1. *J. Mol. Biol.* **1997**, *268*, 539–556. [CrossRef] [PubMed]
69. Bingham, E.; Mannila, H. Random projection in dimensionality reduction: Applications to image and text data. In Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 26–29 August 2001; pp. 245–250.