International Journal of
Molecular Sciences

MDPI

Article

PCLPred: A Bioinformatics Method for Predicting Protein–Protein Interactions by Combining Relevance Vector Machine Model with Low-Rank Matrix Approximation

Li-Ping Li [1,†], Yan-Bin Wang [2,†], Zhu-Hong You [1,*], Yang Li [1] and Ji-Yong An [3]

[1] Department of Information Engineering, Xijing University, Xi'an 710123, China; cs2bioinformatics@gmail.com (L.-P.L.); sxxyliyang@163.com(Y.L.)
[2] Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Science, Urumqi 830011, China; wangyanbin15@mails.ucas.ac.cn
[3] School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 21116, China; ajy@cumt.edu.cn
* Correspondence: zhuhongyou@ms.xjb.ac.cn; Tel.: +86-188-2957-9063
† These authors contributed equally to this work.

Received: 2 March 2018; Accepted: 21 March 2018; Published: 29 March 2018

Abstract: Protein–protein interactions (PPI) are key to protein functions and regulations within the cell cycle, DNA replication, and cellular signaling. Therefore, detecting whether a pair of proteins interact is of great importance for the study of molecular biology. As researchers have become aware of the importance of computational methods in predicting PPIs, many techniques have been developed for performing this task computationally. However, there are few technologies that really meet the needs of their users. In this paper, we develop a novel and efficient sequence-based method for predicting PPIs. The evolutionary features are extracted from the position-specific scoring matrix (PSSM) of protein. The features are then fed into a robust relevance vector machine (RVM) classifier to distinguish between the interacting and non-interacting protein pairs. In order to verify the performance of our method, five-fold cross-validation tests are performed on the *Saccharomyces cerevisiae* dataset. A high accuracy of 94.56%, with 94.79% sensitivity at 94.36% precision, was obtained. The experimental results illustrated that the proposed approach can extract the most significant features from each protein sequence and can be a bright and meaningful tool for the research of proteomics.

Keywords: protein–protein interactions (PPI); low rank; protein sequence; relevance vector machine (RVM); evolutionary information

1. Introduction

Protein–protein interactions (PPI) are a key step in the realization of protein function within cell cycle progression, DNA replication, and signal transmission [1–3]. With the development of high-throughput biological technologies, including a yeast two-hybrid screen (Y2H) [4], protein chip technology [5], mass spectrometry [6], and tandem affinity purification tagging (TAP) [7], more PPI data have been accumulated [8]. PPI datasets have been stored in a number of constructed databases, such as the Molecular Interaction database (MINT), the Database of Interacting Proteins (DIP), and the Biomolecular Interaction Network Database (BIND) [8–10]. However, experimental methods are labor-intensive and time-consuming. The number of PPIs that are validated by these methods represents only a small portion of the entire PPI network. Moreover, the experimental methods

are usually associated with a high rate of both false negative and false positive predictions. All of these drawbacks encourage further research into a computational approach for identifying PPIs.

Different kinds of available protein data are obtained by previous experimental methods, such as the primary, secondary, and tertiary structure of proteins. In order to utilize this wealth of protein data, numerous machine learning approaches have been designed to infer new PPIs. It is popular, among these approaches, to predict PPIs based on the structure of the protein information. For example, Agrawal et al. [11] proposed a computational tool—named a spatial interaction map (SIM)—that utilizes the structure of unbound proteins to detect the residues from PPIs. Qiu et al. [12] presented a novel residue characterization model, based on 3D structures, for the purpose of detecting PPIs. These computational methods—based on structural data—identify the interaction domain by analyzing the hydrophobicity, solvation, protrusion, and accessibility of residues. Since the volume of newly discovered protein sequence data is increasing exponentially, there is an increasingly larger gap between the volume of complex protein structure data, and that of protein sequence data [13,14]. Predicting PPIs based on structure data does not satisfy the requests of the many biochemists who have the sequences, but no structural data. Therefore, it is more important to develop effective computational models based on protein sequence data.

Currently, there are a number of different computational methods designed to implement this pattern in PPI prediction [15–23]. The common computational models for PPI prediction are composed of two key parts, namely, protein feature representation and sample classification. The purpose of the first step is to represent the proteins with useful attributes and transform the samples into feature vectors that are the same size as the sample classifier's inputs. Effective feature descriptors can play an important role in improving the prediction performance of the system.

Previous studies have shown that the evolutionary information on proteins may play a crucial role in predicting PPIs [24,25]. However, it is not easy to include evolutionary information in a protein sequence [26–28]. There is currently no single protein presentation method that takes full advantage of protein evolutionary information. Additionally, sequence evolution information is more difficult to use because of the differences in protein sequence length. In the face of such difficulties, how do we design a way to use the evolution information of proteins to implement the prediction of PPIs efficiently? In order to overcome this problem, we proposed a novel scheme that uses a position-specific scoring matrix (PSSM) to translate the protein sequence into a matrix, in which both the evolutionary information and the amino acid composition are included. Following this, we introduced a low-rank approximation (LRA) method to find the lowest level representation of all of the candidates and accurately recover the row space of the data to achieve high precision.

With regards to the second issue, some machine learning algorithms—such as random forests, neural networks, ensemble classifiers, random projections, and Naïve Bayes classifiers—are proposed for detecting PPIs to improve the accuracy of the prediction model [29–31]. The main trend in computational PPI detection is to achieve the highest precision, rather than speed, in the training of the classified model. Recently, relevance vector machines (RVMs) are a new statistical learning technique that provide the output of the probability classification, which uses Bayesian inferences to obtain a concise solution for the regression and classification [32]. Unlike support vector machines (SVM), RVM classifiers—with fewer input variables—provide better classification estimates for small, high dimensional datasets [33]. In this paper, the performances of RVMs and SVMs for classifying PPIs were compared. Using the PPI dataset, we show that the proposed method can quickly and effectively differentiate interactive protein pairs from large-scale data. The results of the experiment indicate that the proposed technique can complement experimental approaches for identifying PPI interactions.

In this paper, we proposed a novel protein representation method using protein evolutionary information. The main improvement was attributed to the use of LRA, a PSSM, and RVMs. In particular, we first used an LRA method on a PSSM that represented protein in a matrix form to obtain the feature vectors of the protein. Following this, the principal component analysis (PCA) method was employed to eliminate some of the noise and reduce the dimensions of the feature vectors. Finally, we used RVM classifiers to carry out the test. The proposed method was performed on the *Yeast* PPI dataset. The experimental results

show that it is superior to SVM-based methods and other excellent technology that has been developed previously. Therefore, this approach is fit for predicting PPIs. Additionally, a user-friendly web server for predicting PPIs, PCLPred, was developed for academic users at http://219.219.62.123:8888/pclpred/.

The rest of this paper is organized as follows: Section 2 introduces the test results obtained from applying the proposed method, the SVM-based method, and several other existing methods. Section 3 describes the proposed approach. Section 4 summarizes the work presented in this paper.

2. Results and Discussion

2.1. Five-Fold Cross-Validation

In this study, five-fold cross-validation methods were utilized to compare the performance of this model with other competing approaches. The whole PPI dataset is randomly divided into five roughly-equivalent subsets, each containing approximately equal amounts of interacting and non-interacting proteins. Four of the subsets are used for training and the remaining one is used for the test. This process is repeated five times, using a different subset of the test each time. The average of the five results is then calculated to ensure the highest level of fairness.

2.2. Comparison with the SVM-Based Approach Using the Same Feature Representation

In order to effectively assess the performance of the SVM classifier, we compared its performance with that of a state-of-the-art SVM classifier with the same feature extraction method on the *Yeast* dataset [34]. The LIBSVM (A Library for Support Vector Machines) tool provides an interface to facilitate the use of the SVM classifier. The cross-validation strategy is employed to optimize the related parameters of the SVM. Consequently, the parameters (c, g) are set to 0.8 and 0.4, respectively. Furthermore, the radial basis function is taken as the kernel function.

The result of applying the two methods to the *Yeast* dataset are presented in Table 1, and the corresponding receiver operating characteristic (ROC) curves are shown in Figure 1. The prediction performance of the SVM classifier can be seen, from Table 1, to have achieved 89.4% accuracy, 88.5% sensitivity, 90.3% specificity, and 81.1% Matthews Correlation Coefficient (MCC). The average prediction results of applying the RVM classifier were 94.6% accuracy (which is 5.2% higher than the SVMs classifier) and 94.8% sensitivity (which is 6.3% higher than SVMs classifier). Several other indicators of the RVM classifier's performance—shown in Table 1—are 4.0% above the performance of the SVM classifier. This comparison proves that the effect of using the RVM classifier to predict PPIs can be clearly distinguished from the effect of using the SVM classifier. Additionally, Figure 1 indicates that the ROC curves of the two classifiers also show that RVM classifier can be more powerful in detecting PPI performance than the SVM classifier.

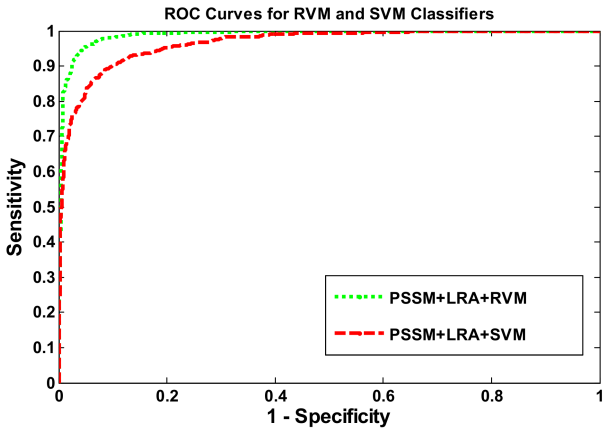

Figure 1. A comparison of the receiver operating characteristic (ROC) curves of the relevance vector machines (RVMs) classifier and the support vector machines (SVMs) classifier on the *Yeast* dataset.

Table 1. Five-fold cross-validation results shown using our proposed method on the *Yeast* dataset.

Model	Testing Set	Accuracy	Sensitivity	Specificity	PPV	NPV	MCC
PSSM+ LR+RVM	1	94.7%	95.4%	94.0%	93.9%	95.46%	89.3%
	2	95.3%	96.1%	94.5%	94.7%	95.96%	91.1%
	3	93.9%	93.9%	93.8%	93.8%	93.91%	88.5%
	4	93.8%	93.6%	94.1%	94.4%	93.22%	88.4%
	5	95.1%	94.9%	95.2%	94.9%	95.2%	90.6%
	Average	**94.6 ± 0.6%**	**94.8 ± 1.0%**	**94.3 ± 0.5%**	**94.3 ± 0.4%**	**94.75 ± 1.1%**	**89.6 ± 1.2%**
PSSM+ LR+SVM	1	88.3%	87.3%	89.3%	88.8%	87.8%	79.4%
	2	89.3%	89.4%	89.1%	89.2%	89.3%	80.8%
	3	89.8%	89.2%	90.3%	90.7%	88.8%	81.6%
	4	89.7%	88.3%	91.2%	90.9%	88.6%	81.6%
	5	90.0%	88.4%	91.5%	90.8%	89.2%	81.9%
	Average	**89.4 ± 0.6%**	**88.5 ± 0.8%**	**90.3 ± 1.0%**	**90.1 ± 1.0%**	**88.7 ± 0.5%**	**81.1 ± 1.0%**

SVM: support vector machine; PSSM: position specific scoring matrix; AB: average blocks; RVM: relevance vector machine; PPV: Positive Predictive Value; NPV: Negative Predictive Value; MCC: Matthews Correlation Coefficient.

The reasons for this method producing better classification results come from the following points: (1) Based on a Bayesian framework to build a learning machine, the RVM classifier is conducive to making more scientific decisions based on the information; (2) in the choice of the kernel function, the RVM classifier is not limited by the Mercer theorem, and can construct any kernel function; (3) there is no need to set penalties. The penalty factor in the SVM classifier is a constant that balances the empirical risk and the confidence interval. The experimental results are very sensitive to the data. An improper setting may cause over-learning and other problems. The parameters in the RVM classifier, however, are automatically assigned; (4) compared to the SVM classifier, the RVM classifier is sparser, which means that the test time is shorter, making it more suitable for online testing. It is well known that the number of SVM support vectors grows linearly with the increase of the training samples, which is obviously not convenient when the training samples are very large. Although the RVM correlation vector also increases with the training samples, the growth rate is much slower than that of the SVM support vectors; and (5) previous research indicates that the RVM classifier has a better generalization performance than the SVM classifier. Additionally, when compared with the SVM classifier, the RVM classifier not only produces a binary output, but also gets the probability of the output.

2.3. A Comparison of the Proposed Method with Other Methods

Currently, many methods that are based on machine learning theory have been proposed for sequences-based PPIs. To assess the ability of the proposed approach, several existing techniques [35] are applied to the *Yeast* dataset and their results are compared to the results of our method. The comparison of the results of these methods is listed in Table 2. Table 2 clearly indicates that the proposed method achieved the highest average accuracy (94.6%) out of all of these methods. At the same time, the sensitivity and precision of the proposed technique are also superior to those of the other techniques. All of these results indicated that the RVM classifier, using the features vector that was extracted by the PSSM, LRA, and the PCA method, can substantially improve the quality of PPI prediction. This is mainly because of the efficient feature extraction strategy and the powerful classifier.

Table 2. The prediction ability of the different methods on the *Yeast* dataset.

Model	Testing Set	Acc (%)	Sen (%)	Pre (%)	Mcc (%)
Guos' work [35]	ACC	89.3 ± 2.6	89.9 ± 3.6	88.8 ± 6.1	N/A
	AC	87.4 ± 1.3	87.3 ± 4.6	87.8 ± 4.3	N/A
Zhous' work [36]	SVM+LD	88.6 ± 0.3	87.4 ± 0.2	89.5 ± 0.6	77.2 ± 0.7
Yangs' work [37]	Cod1	75.1 ± 1.1	75.8 ± 1.2	74.8 ± 1.2	N/A
	Cod2	80.0 ± 1.0	76.8 ± 0.6	82.2 ± 1.3	N/A
	Cod3	80.4 ± 0.4	78.1 ± 0.9	81.7 ± 0.9	N/A
	Cod4	86.2 ± 1.1	81.0 ± 1.7	90.2 ± 1.3	N/A
Yous' work [38]	PCA-EELM	87.0 ± 0.2	86.2 ± 0.4	87.6 ± 0.3	77.4 ± 0.4
Proposed method	LRA+RVM	94.6 ± 0.6	94.8 ± 1.0	94.4 ± 0.4	89.6 ± 1.2

ACC: Auto Covariance; LD: Local Description; PCA: Principal Component Analysis; EELM: Ensemble Extreme Learning Machines; N/A: Not Available; Acc: Accuracy; Sen: sensitivity; Pre: precision; Mcc: Matthew's Correlation Coefficient.

2.4. An Assessment of the Prediction Performance on the Helicobacter pylori PPI Dataset

In order to further investigate the prediction performance of our approach, we also compared the proposed approach with several other existing methods on the *Helicobacter pylori* PPI dataset. The prediction results for the abovementioned methods are reported in Table 3. In order to achieve a fair measure of randomness, we calculated the average of the measure values over five runs. We can observe from Table 3 that this method can achieve a good result, with 84.7% accuracy, 85.9% precision, and 84.4% sensitivity. It should be noticed that the precision and accuracy achieved by the proposed method are superior to those of the other methods.

Table 3. The prediction ability of the different methods on the *Helicobacter pylori* protein–protein interactions (PPIs) dataset.

Methods	Acc(%)	Sen (%)	Pre (%)	Mcc (%)
HKNN	84.0	86.0	84.0	N/A
Phylogenetic bootstrap	75.8	69.8	80.2	N/A
Signature Products	83.4	79.9	85.7	N/A
Boosting	79.5	80.4	81.7	N/A
Proposed method	**84.7 ± 1.0**	**84.4 ± 1.2**	**85.9 ± 0.8**	**76.7 ± 1.0**

HKNN: Hyperplane Distance Nearest Neighbor.

3. Materials and Methods

3.1. Dataset

In this paper, the proposed approach was verified on the high-confidence *Yeast* and *Helicobacter pylori* PPI datasets. We gathered the *Yeast* dataset from the publicly available Database of Interacting Proteins (DIP) [8]. For the purpose of ensuring the effectiveness of the experiment, we removed the protein pairs of less than fifty residues and greater than 40% sequence identity. By performing this screening work, the remaining 5594 protein pairs are reserved for building the positive dataset. The additional 5594 non-interacting protein pairs, with different subcellular localizations, were then used to build the negative dataset. As a result, the whole *Yeast* dataset finally consisted of 11,188 protein pairs. In order to further verify the general applicability of the proposed method, we also evaluated our method on the *Helicobacter pylori* PPI dataset. In total, we obtained 1458 positive samples and 1458 negative samples, as described by Martin et al. [39].

3.2. Position Specific Scoring Matrix (PSSM)

PSSM is a type of scoring matrix that was proposed by Gribskov et al. [24]. It is used to perform BLAST (Basic Local Alignment Search Tool) searches, where amino acid substitution scores are assigned to a specific location in the proteins' multiple sequence alignments. It has been successfully applied in various fields of biological information because it contains the evolutionary information of proteins. PSSM is represented as a $T \times 20$ matrix that can be interpreted as $M = \{c_{i,j} : i = 1 \cdots T \text{ and } j = 1 \cdots 20\}$. The representation of PSSM is as follows:

$$M = \begin{bmatrix} c_{1,1} & c_{1,2} & \cdots & c_{1,20} \\ c_{2,1} & c_{2,2} & \cdots & c_{2,20} \\ \vdots & \vdots & \vdots & \vdots \\ c_{L,1} & c_{L,2} & \cdots & c_{L,20} \end{bmatrix} \quad (1)$$

The elements in this matrix are generally expressed as integers (negative or positive). A higher score indicates that a given amino acid substitution occurs frequently in the alignment, while a lower score indicates a lower frequency of the substitution.

We created the PSSM using a Position-Specific Iterated BLAST (PSI-BLAST, Bethesda, MD, USA), which found a protein sequence that was similar to the query sequence, and then constructed the PSSM from the obtained alignment. In this work, we set the number of iterations to three and the *e*-value to 0.001 and *t*, respectively, in order to obtain a highly broad homologous sequence.

3.3. Low-Rank Approximation (LRA)

LRA is a widely used method for matrix analysis, where the cost function measures the fit between an approximation matrix (optimization variable) and a given sparse matrix, constrained by the reduced rank of the approximation matrix [40,41]. In this case, using LRA on the PSSM of the obtained protein sequences results in a descriptor containing evolutionary information that is used for representing a protein. For a $20 \times L$ feature matrix N, the LRA would be written as follows:

$$\min_{\hat{N}} \|N - \hat{N}\|_F \quad (2)$$

$$\text{Subject to : } rank(\hat{N}) \leq r \quad (3)$$

where $\|\bullet\|_F$ represents the Frobenius norm. Formula (2) is solved using the singular value decomposition (SVD) method.

Let $N = U \sum V^T \in R^{m \times n}$ be the SVD of N and partition U, $\sum =: diag(\sigma_1, \sigma_2, \sigma_3, \ldots, \sigma_{20})$, and V as follows:

$$U =: \begin{bmatrix} U_1 & U_2 \end{bmatrix}, \sum =: \begin{bmatrix} \sum_1 & 0 \\ 0 & \sum_2 \end{bmatrix}, \text{and } V =: \begin{bmatrix} V_1 & V_2 \end{bmatrix} \quad (4)$$

where $\sum_1$ is a square array of r. U_1 and V_1 represent different matrices, and their sizes are $m \times r$ and $n \times r$. The rank-r matrix can then be gained as follows:

$$\hat{N}^* = U_1 \sum\nolimits_1 {V_1}^T \quad (5)$$

where $\|N - \hat{N}^*\|_F = \min_{rank(\hat{N}) \leq r} \|N - \hat{N}\|_F = \sqrt{\sigma_{r+1}^2 + \sigma_{r+2}^2 + \ldots + \sigma_m^2}$.

The ${\Sigma_1}^{1/2}$, with dimensions *r*-by-*r*, can be obtained by computing the square root of the reduced matrix Σ_1, in which the sequence order information of the protein is contained. It is noteworthy that the feature matrix N of the protein may have a different number of columns, which is caused by the unequal lengths of protein sequences. However, the $U_1 {\Sigma_1}^{1/2}$ is a fixed length (a $20 \times r$ matrix).

We form a vector from the gained matrix $U_1{\Sigma_1}^{1/2}$ by concatenating all of the rows, from row 1 to 20, of matrix $U_1{\Sigma_1}^{1/2}$. Therefore, the feature descriptor consists of a total of $20 \times r$ descriptor values. Considering the trade-off between the cost of computing for extracting the protein feature and the overall prediction accuracy, the optimal rank is 5. We connect the descriptors of the two protein sequences to represent an interaction pair.

3.4. Properties of the Proposed Algorithm

Based on orthogonal triangular decomposition theory and LRA theory, the properties of the PSSM feature extraction algorithm are deduced.

Lemma 1. *Suppose that matrix $\hat{N}^*$ in (5) satisfies (3). For the Frobenius norm, if $r \leq rank(N)$, then $\hat{N}^*$ is unique if and only if N's rth and $(r+1)$th largest singular values differ.*

Proof. $\hat{N}^*$ is a solution to

$$\hat{N}^* := \underset{\hat{M}}{\operatorname{argmin}} \|N - \hat{N}\|_F \tag{6}$$

$$s.t. rank(\hat{N}) \leq r.$$

and $\hat{N}^* := U^* \sum^* (V^*)^T$ is an SVD of $\hat{N}^*$. Based on the single invariance of the Frobenius norm, we have

$$\|N - \hat{N}^*\|_F = \|(U^*)^T (N - \hat{N}^*) V^*\|_F = \|(U^*)^T N V^* - \sum^*\|_F \tag{7}$$

where $(U^*)^T N V^* = \hat{N}$. Partition

$$\hat{N} = \begin{bmatrix} \hat{N}_{11} & \hat{N}_{12} \\ \hat{N}_{21} & \hat{N}_{22} \end{bmatrix} \tag{8}$$

conformably with $\sum^* = \begin{bmatrix} \sum^*{}_1 & 0 \\ 0 & 0 \end{bmatrix}$ and observe that

$$rank\left(\begin{bmatrix} \sum^*{}_1 & \hat{N}_{12} \\ 0 & 0 \end{bmatrix}\right) \leq m \text{and} \hat{N}_{12} \neq 0 \Rightarrow \|\hat{N} - \begin{bmatrix} \sum^*{}_1 & \hat{N}_{12} \\ 0 & 0 \end{bmatrix}\|_F < \|\hat{N} - \begin{bmatrix} \sum^*{}_1 & 0 \\ 0 & 0 \end{bmatrix}\|_F \tag{9}$$

Thus, $\hat{N}_{12} = 0$. Similarly, $\hat{N}_{21} = 0$. Observe also that

$$rank\left(\begin{bmatrix} \hat{N}_{11} & 0 \\ 0 & 0 \end{bmatrix}\right) \leq m \text{and } \hat{N}_{11} \neq \sum_1{}^* \Rightarrow \|\hat{N} - \begin{bmatrix} \hat{N}_{11} & 0 \\ 0 & 0 \end{bmatrix}\|_F < \|\hat{N} - \begin{bmatrix} \sum^*{}_1 & 0 \\ 0 & 0 \end{bmatrix}\|_F \tag{10}$$

Thus, $\hat{N}_{11} = \sum_1^*$. Therefore,

$$\hat{N} = \begin{bmatrix} \sum^*{}_1 & 0 \\ 0 & \hat{N}_{22} \end{bmatrix} \tag{11}$$

Let $\hat{N}_{22} = U_{22} \sum_{22} {V_{22}}^T$ be the SVD of $\hat{N}_{22}$. Then the matrix

$$\begin{bmatrix} I & 0 \\ 0 & {U_{22}}^T \end{bmatrix} \hat{N} \begin{bmatrix} I & 0 \\ 0 & V_{22} \end{bmatrix} = \begin{bmatrix} \sum_1^* & 0 \\ 0 & \sum_{22} \end{bmatrix} \tag{12}$$

has the optimal rank-m approximation $\sum^* = \begin{bmatrix} \sum_1^* & 0 \\ 0 & 0 \end{bmatrix}$, such that

$$\min(diag(\textstyle\sum_1^*)) > \max(diag(\sum_{22})) \tag{13}$$

Therefore,

$N = U^* \begin{bmatrix} I & 0 \\ 0 & U_{22} \end{bmatrix} \begin{bmatrix} \sum_1^* & 0 \\ 0 & \sum_{22} \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & {V_{22}}^T \end{bmatrix} (V^*)^T$ is an SVD of N.
Thus, if $\sigma_m > \sigma_{m+1}$, the rank-m truncated SVD

$$\hat{N}^* = U^* \begin{bmatrix} \sum_1^* & 0 \\ 0 & \sum_{22} \end{bmatrix} (V^*)^T = U^* \begin{bmatrix} I & 0 \\ 0 & U_{22} \end{bmatrix} \begin{bmatrix} \sum_1^* & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & {V_{22}}^T \end{bmatrix} (V^*)^T \tag{14}$$

is unique and $\hat{N}^*$ is the unique solution of LRA. □

The salient feature of Lemma 1 is that, although the rank constraint is highly non-convex and non-linear, one is still able to efficiently solve (2) using the SVD method. Additionally, under all of the consistent rules, there is an optimal solution under the *Frobenius* norm.

3.5. Relevance Vector Machine (RVM) Model

The RVM model is a probabilistic model under a Bayesian framework, developed by Tipping et al. [32,33,42]. It has been widely applied for solving classification and regression problems. Assuming that the training datasets are $(x_n,y_n)_{n=1}{}^N$ for binary classification problems, $x_n \in R^d$ is the training sample; $t_n \in (0,1)$ denotes the label of the training dataset; t_i is the label of the testing dataset; $t_i = b_i + \varepsilon_i$, where $b_i = w^T\varphi(x_i) = \sum_{j=1}^N w_j K(x_i, x_j) + w_0$, is the classification model; and ε_i is the additional noise, with a variance of σ^2 and a mean value of zero, where $\varepsilon_i \sim N(0,\sigma^2), y_i \sim N(b_i,\sigma^2)$. The training datasets are assumed to be independent and distributed identically. The observation of vector t follows the distribution as follows:

$$m\left(y\middle|x,c,\sigma^2\right) = \left(2\pi\sigma^2\right)^{-N/2} \exp[-\frac{1}{2\sigma^2}||y-\partial c||^2] \tag{15}$$

where ∂ meets the following definition:

$$\partial = \begin{pmatrix} 1 & k(x_1,x_1)\cdots & k(x_1,x_N)) \\ \cdots & \cdots & \cdots \\ 1 & k(x_N,x_1)\ldots & k(x,x_N) \end{pmatrix} \tag{16}$$

The method used by the RVMs to predict the label t * of a test sample is given by:

$$m(y_*|y) = \int p\left(y_*\middle|c,\sigma^2\right) p\left(c,\sigma^2\middle|y\right) dwd\sigma^2 \tag{17}$$

In order to reduce the computational complexity of the kernel function and ensure that the majority of the weight vector has a value of zero, the weight vector w is limited by extra conditions. Assuming that $c_i \sim N\left(0,\ x_i^{-1}\right)$, $p(w|x) = \prod_{i=0}^N p(c_i|x_i)$, where x is a hyper-parameters vector.

$$m(t_*|t) = \int p\left(t_*\middle|w,x,\sigma^2\right) p\left(c,x,\sigma^2\middle|t\right) dwdad\sigma^2 \tag{18}$$

$$m\left(t_*\middle|c,x,\sigma^2\right) = N\left(t_*\ \middle|b(a_*;c),\sigma^2\right) \tag{19}$$

We get $p\left(c,x,\sigma^2\middle|t\right)$ using the Bayesian formula

$$m\left(c,x,\sigma^2\middle|t\right) = p(x,\sigma^2\middle|t)p\left(x\middle|x,\sigma^2,t\right) \tag{20}$$

$$m\left(c\middle|x,\sigma^2,t\right) = p\left(t\middle|c,\sigma^2\right)p(c|x)/p\left(t\middle|x,\sigma^2\right) \tag{21}$$

The integral of the product of $p(t|x,\sigma^2)$ and $p(c|x)$ is given by

$$m\left(t\middle|x,\sigma^2\right) = (2\pi)^{-N/2}|\Omega|^{-1/2}\exp\left(-\frac{t^T\Omega^{-1}t}{2}\right) \tag{22}$$

$$\Omega = \sigma^2 I + \partial A^{-1}\partial^T,\ A = diag(x_0, x_1, \ldots, x_N) \tag{23}$$

$$m\left(c\middle|x,\sigma^2,t\right) = (2\pi)^{-(N+1)/2}|\Sigma|^{-1/2}\exp\left(-\frac{(c-u)^T(c-u)}{2}\right) \tag{24}$$

$$\Sigma = \left(\sigma^{-2}\partial^T\partial + A\right)^{-1} \tag{25}$$

$$u = \sigma^{-2}\Sigma\partial^T t \tag{26}$$

The maximum likelihood method was used to solve $m(x,\sigma^2|t) \propto m\left(t|x,\sigma^2\right)m(x)m\left(\sigma^2\right)$ and $m(x,\sigma^2|t)$, and is represented by

$$\left(x_{MP},\sigma^2_{MP}\right) = \underset{x,\sigma^2}{\arg maxp}(t|x,\sigma^2) \tag{27}$$

The iterative process of x_{MP} and σ^2_{MP} is as follows:

$$\begin{cases} x_i^{new} = \frac{\gamma_i}{\mu_i^2} \\ (\sigma^2)^{new} = \frac{||t-\partial\mu||^2}{N-\sum_{i=0}^{N}\mu_i} \\ \gamma_i = 1 - x_i\sum i,i \end{cases} \tag{28}$$

where $\sum i,i$ represents the ith element on the diagonal of Σ, and the initial value of a and σ^2 are determined via the approximation of a_{MP} and σ^2_{MP}, by continuously using Formula (19).

3.6. Procedure of the Proposed Method

In the study, the workflow of the PCLPred method is presented in Figure 2. More specifically, the protein amino acids sequence datasets are downloaded from DIP. The CD-HIT (Cluster Database at High Identity with Tolerance) and PSI-BLAST programs are then used to remove sequence redundancy and generate PSSM, respectively [43]. Following this, LRA is employed to obtain the feature representation from PSSM, which contains a large volume of valuable evolutionary knowledge for PPI prediction. After the dimensionality reduction—using the PCA technique—the significant features are extracted and used as input features to train the RVM classifier. Finally, the prediction performance is evaluated using five-fold cross-validations [44–48].

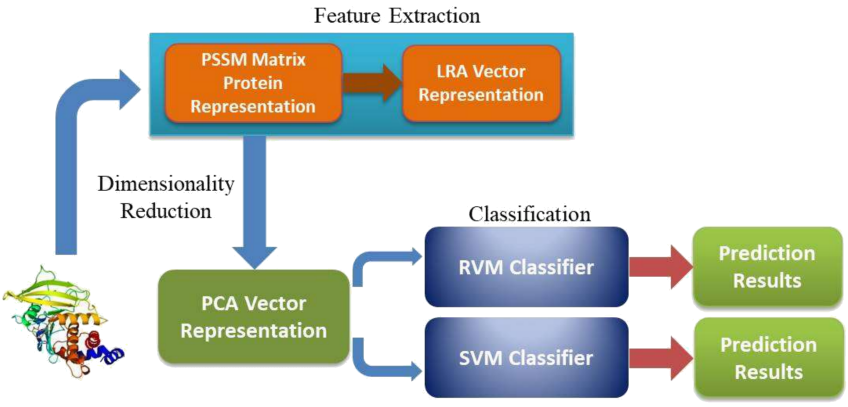

Figure 2. The flow chart of the proposed method.

3.7. Performance Evaluation

In order to evaluate the performance of the designed model, a number of validation measures are employed.

(1) Overall prediction accuracy:

$$Accuracy = \frac{T_P + T_N}{T_P + F_P + T_N + F_N} \tag{29}$$

(2) Sensitivity:

$$Sensitivity = \frac{T_P}{T_P + F_N} \tag{30}$$

(3) Specificity:

$$Specificity = \frac{T_N}{T_N + F_P} \tag{31}$$

(4) Positive predictive value:

$$PPV = \frac{T_P}{T_P + F_P} \tag{32}$$

(5) Negative predictive value:

$$NPV = \frac{T_N}{T_N + F_N} \tag{33}$$

(6) *F*-score:

$$F_s = 2 \times \frac{Sen \times PPV}{Sen + PPV} \tag{34}$$

(7) Matthews correlation coefficient:

$$MCC = \frac{T_P \times T_N - F_P \times F_N}{\sqrt{(T_P + F_N) \times (T_N + F_P) \times (T_P + F_P) \times (T_N + F_N)}} \tag{35}$$

where T_P is true positive, indicating that the total number of interactive proteins will be predicted correctly; F_P is false positive, indicating the total number of these proteins pairs that have no interaction, but are determined as interacting; F_N is false negative, indicating the total number of interactive proteins that are determined as non-interacting; and T_N is true negative, indicating the total number of these proteins pairs that have no interaction that are determined correctly. Additionally, the ROC curve is adopted as a measure that is used to evaluate the prediction performance of the different methods [49,50].

4. Conclusions

In this study, we proposed a novel computation-based automated decision-making method by employing the RVM model combined with the LRA method and PSSM. More specifically, LRA is employed to obtain the feature representation from PSSM, which contains a large volume of valuable evolutionary knowledge for PPI prediction. The RVM classifier is then applied to predict novel PPIs. Extensive computational experiments are performed on several PPI datasets in order to evaluate the PPI identification ability of the developed approach. These experimental results have proven that the PPI identification ability of this approach is clearly stronger than that of the SVM-based method and several other existing approaches. The promising results demonstrate that the proposed method is an efficient and reliable approach to detecting PPIs. It is also a practical tool that will help to advance research in the field of bioinformatics.

Acknowledgments: This study is supported by the National Science Foundation of China and the Pioneer Hundred Talents Program of Chinese Academy of Sciences, under Grants 61572506.

Author Contributions: Li-Ping Li, Yan-Bin Wang, Zhu-Hong You prepared the data sets, conceived the algorithm, carried out the analyses, carried out experiments, and Yang Li and Ji-Yong An wrote the manuscript. All authors read and approved the final manuscript.

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

References

1. Uetz, P.; Giot, L.; Cagney, G.; Mansfield, T.A.; Judson, R.S.; Knight, J.R.; Lockshon, D.; Narayan, V.; Srinivasan, M.; Pochart, P.; et al. A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. *Nature* **2000**, *403*, 623–627. [CrossRef] [PubMed]
2. Zhang, Q.C.; Petrey, D.; Deng, L.; Qiang, L.; Shi, Y.; Thu, C.A.; Bisikirska, B.; Lefebvre, C.; Accili, D.; Hunter, T. Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature* **2012**, *7421*, 556–560. [CrossRef] [PubMed]
3. Gavin, A.C.; Bosche, M.; Krause, R.; Grandi, P.; Marzioch, M.; Bauer, A.; Schultz, J.; Rick, J.M.; Michon, A.M.; Cruciat, C.M.; et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **2002**, *415*, 141–147. [CrossRef] [PubMed]
4. Ito, T.; Chiba, T.; Ozawa, R.; Yoshida, M.; Hattori, M.; Sakaki, Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 4569–4574. [CrossRef] [PubMed]
5. Zhu, H.; Bilgin, M.; Bangham, R.; Hall, D.; Casamayor, A.; Bertone, P.; Lan, N.; Jansen, R.; Bidlingmaier, S.; Houfek, T. Global analysis of protein activities using proteome chips. *Science* **2001**, *293*, 2101–2105. [CrossRef]
6. Nesvizhskii, A.I.; Keller, A.; Kolker, E.; Aebersold, R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* **2003**, *75*, 4646–4658. [CrossRef] [PubMed]
7. Puig, O.; Caspary, F.; Rigaut, G.; Rutz, B.; Bouveret, E.; Bragado-Nilsson, E.; Wilm, M.; Séraphin, B. The tandem affinity purification (tap) method: A general procedure of protein complex purification. *Methods* **2001**, *24*, 218–229. [CrossRef] [PubMed]
8. Xenarios, I.; Salwínski, L.; Duan, X.J.; Higney, P.; Kim, S.M.; Eisenberg, D. Dip, the database of interacting proteins: A research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.* **2002**, *30*, 303–305. [CrossRef] [PubMed]
9. Chatr-Aryamontri, A.; Ceol, A.; Palazzi, L.M.; Nardelli, G.; Schneider, M.V.; Castagnoli, L.; Cesareni, G. Mint: The molecular interaction database. *Nucleic Acids Res.* **2010**, *40*, D572–D574. [CrossRef] [PubMed]
10. Bader, G.D.; Donaldson, I.; Wolting, C.; Ouellette, B.F.F.; Pawson, T.; Hogue, C.W.V. Bind—The biomolecular interaction network database. *Nucleic Acids Res.* **2001**, *29*, 242–250. [CrossRef] [PubMed]
11. Agrawal, N.J.; Helk, B.; Trout, B.L. *A Computational Tool to Predict the Evolutionarily Conserved Protein-Protein Interaction Hot-Spot Residues from the Structure of the Unbound Protein*; Asia Publishing Houser: Cambridge, MA, USA, 2014; pp. 326–333.
12. Qiu, Z.; Wang, X. Prediction of protein-protein interaction sites using patch-based residue characterization. *J. Theor. Biol.* **2012**, *293*, 143–150. [CrossRef] [PubMed]
13. Liu, B.; Xu, J.; Zou, Q.; Xu, R.; Wang, X.; Chen, Q. Using distances between top-n-gram and residue pairs for protein remote homology detection. *BMC Bioinform.* **2014**, *15*, S3. [CrossRef] [PubMed]
14. Liu, B.; Wang, X.; Zou, Q.; Dong, Q.; Chen, Q. Protein remote homology detection by combining chou's pseudo amino acid composition and profile-based protein representation. *Mol. Inf.* **2013**, *32*, 775–782. [CrossRef] [PubMed]
15. Zhu, L.; You, Z.H.; Huang, D.S. Increasing the reliability of protein–protein interaction networks via non-convex semantic embedding. *Neurocomputing* **2013**, *121*, 99–107. [CrossRef]
16. Huang, Q.; You, Z.; Zhang, X.; Zhou, Y. Prediction of protein–protein interactions with clustered amino acids and weighted sparse representation. *Int. J. Mol. Sci.* **2015**, *16*, 10855–10869. [CrossRef] [PubMed]
17. Huang, Y.-A.; You, Z.-H.; Gao, X.; Wong, L.; Wang, L. Using weighted sparse representation model combined with discrete cosine transformation to predict protein-protein interactions from protein sequence. *BioMed Res. Int.* **2015**, *2015*, 902198. [CrossRef] [PubMed]
18. Lei, Y.-K.; You, Z.-H.; Ji, Z.; Zhu, L.; Huang, D.-S. Assessing and predicting protein interactions by combining manifold embedding with multiple information integration. *BMC Bioinform.* **2012**, *13*, S3. [CrossRef] [PubMed]

19. Luo, X.; Ming, Z.; You, Z.; Li, S.; Xia, Y.; Leung, H. Improving network topology-based protein interactome mapping via collaborative filtering. *Knowl. Based Syst.* **2015**, *90*, 23–32. [CrossRef]
20. 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] [PubMed]
21. You, Z.-H.; Chan, K.C.; Hu, P. Predicting protein-protein interactions from primary protein sequences using a novel multi-scale local feature representation scheme and the random forest. *PLoS ONE* **2015**, *10*, e0125811. [CrossRef] [PubMed]
22. 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] [PubMed]
23. Zhu, L.; You, Z.-H.; Huang, D.-S.; Wang, B. T-lse: A novel robust geometric approach for modeling protein-protein interaction networks. *PLoS ONE* **2013**, *8*, e58368. [CrossRef] [PubMed]
24. Ahmad, S.; Sarai, A. Pssm-based prediction of DNA binding sites in proteins. *BMC Bioinform.* **2005**, *6*, 33. [CrossRef] [PubMed]
25. Li, Z.-W.; You, Z.-H.; Chen, X.; Li, L.-P.; Huang, D.-S.; Yan, G.-Y.; Nie, R.; Huang, Y.-A. Accurate prediction of protein-protein interactions by integrating potential evolutionary information embedded in pssm profile and discriminative vector machine classifier. *Oncotarget* **2017**, *8*, 23638. [CrossRef] [PubMed]
26. Nanni, L. Hyperplanes for predicting protein–protein interactions. *Neurocomputing* **2005**, *69*, 257–263. [CrossRef]
27. Nanni, L. Fusion of classifiers for predicting protein–protein interactions. *Neurocomputing* **2005**, *68*, 289–296. [CrossRef]
28. Nanni, L. An ensemble of k-local hyperplanes for predicting protein-protein interactions. *Neurocomputing* **2006**, *22*, 1207–1210. [CrossRef] [PubMed]
29. Wei, L.; Xing, P.; Shi, G.; Ji, Z.L.; Zou, Q. Fast prediction of protein methylation sites using a sequence-based feature selection technique. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2017**. [CrossRef] [PubMed]
30. Wei, L.; Tang, J.; Zou, Q. Local-dpp: An improved DNA-binding protein prediction method by exploring local evolutionary information. *Inf. Sci.* **2016**, *384*, 135–144. [CrossRef]
31. 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]
32. Wei, L.; Yang, Y.; Nishikawa, R.M.; Wernick, M.N.; Edwards, A. Relevance vector machine for automatic detection of clustered microcalcifications. *IEEE Trans. Med. Imaging* **2005**, *24*, 1278–1285. [PubMed]
33. Widodo, A.; Kim, E.Y.; Son, J.D.; Yang, B.S.; Tan, A.C.C.; Gu, D.S.; Choi, B.K.; Mathew, J. Fault diagnosis of low speed bearing based on relevance vector machine and support vector machine. *Expert Syst. Appl.* **2009**, *36*, 7252–7261. [CrossRef]
34. Chang, C.C.; Lin, C.J. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2001**, *2*, 1–27. [CrossRef]
35. Yanzhi, G.; Lezheng, Y.; Zhining, W.; Menglong, L. Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences. *Nucleic Acids Res.* **2008**, *36*, 3025–3030.
36. Zhou, Y.Z.; Gao, Y.; Zheng, Y.Y. *Prediction of Protein-Protein Interactions Using Local Description of Amino Acid Sequence*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 254–262.
37. Lei, Y.; Jun-Feng, X.; Jie, G. Prediction of protein-protein interactions from protein sequence using local descriptors. *Protein Pept. Lett.* **2010**, *17*, 1085–1090.
38. You, Z.H.; Lei, Y.K.; Zhu, L.; Xia, J.; Wang, B. Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis. *BMC Bioinform.* **2013**, *14*, S10. [CrossRef] [PubMed]
39. Martin, S.; Roe, D.; Faulon, J.L. Predicting protein-protein interactions using signature products. *Bioinformatics* **2005**, *21*, 218–226. [CrossRef] [PubMed]
40. Liberty, E.; Woolfe, F.; Martinsson, P.G.; Rokhlin, V.; Tygert, M. Randomized algorithms for the low-rank approximation of matrices. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 20167–20172. [CrossRef] [PubMed]
41. Markovsky, I. Structured low-rank approximation and its applications. *Automatica* **2008**, *44*, 891–909. [CrossRef]

42. Tipping, M.E. Sparse bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.* **2001**, *1*, 211–244.
43. Zou, Q.; Hu, Q.; Guo, M.; Wang, G. Halign: Fast multiple similar DNA/RNA sequence alignment based on the centre star strategy. *Bioinformatics* **2015**, *31*, 2475–2481. [CrossRef] [PubMed]
44. Zou, Q.; Zeng, J.; Cao, L.; Ji, R. A novel features ranking metric with application to scalable visual and bioinformatics data classification. *Neurocomputing* **2016**, *173*, 346–354. [CrossRef]
45. Zou, Q.; Ju, Y.; Li, D. Protein Folds Prediction with Hierarchical Structured SVM. *Curr. Proteom.* **2016**, *13*, 79–85.
46. Zou, Q. Editorial (Thematic Issue: Machine Learning Techniques for Protein Structure, Genomics Function Analysis and Disease Prediction). *Curr. Proteom.* **2016**, *13*, 77–78. [CrossRef]
47. Zou, Q.; Xie, S.; Lin, Z.; Wu, M.; Ju, Y. Finding the Best Classification Threshold in Imbalanced Classification. *Big Data Res.* **2016**, *5*, 2–8. [CrossRef]
48. 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] [PubMed]
49. Wang, Y.; You, Z.; Xiao, L.; Xing, C.; 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]
50. 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]