Int. J. Mol. Sci. **2014**, *15*, 3220-3233; doi:10.3390/ijms15023220

OPEN ACCESS

International Journal of

Molecular Sciences

ISSN 1422-0067

www.mdpi.com/journal/ijms

Article

Prediction of Protein–Protein Interaction with Pairwise Kernel Support Vector Machine

Shao-Wu Zhang [1,2,]*, Li-Yang Hao [1] and Ting-He Zhang [1]

[1] College of Automation, Northwestern Polytechnical University, Xi'an 710072, China;
E-Mails: haoliyang0306@sina.com (L.-Y.H.); zth86z68@163.com (T.-H.Z.)

[2] Key Laboratory of Information Fusion Technology, Ministry of Education, Xi'an 710072, China

* Author to whom correspondence should be addressed; E-Mail: zhangsw@nwpu.edu.cn;
Tel.: +86-29-8843-1308; Fax: +86-29-8843-1306.

Received: 1 January 2014; in revised form: 27 January 2014 / Accepted: 29 January 2014 / Published: 21 February 2014

Abstract: Protein–protein interactions (PPIs) play a key role in many cellular processes. Unfortunately, the experimental methods currently used to identify PPIs are both time-consuming and expensive. These obstacles could be overcome by developing computational approaches to predict PPIs. Here, we report two methods of amino acids feature extraction: (i) distance frequency with PCA reducing the dimension (DFPCA) and (ii) amino acid index distribution (AAID) representing the protein sequences. In order to obtain the most robust and reliable results for PPI prediction, pairwise kernel function and support vector machines (SVM) were employed to avoid the concatenation order of two feature vectors generated with two proteins. The highest prediction accuracies of AAID and DFPCA were 94% and 93.96%, respectively, using the 10 CV test, and the results of pairwise radial basis kernel function are considerably improved over those based on radial basis kernel function. Overall, the PPI prediction tool, termed PPI-PKSVM, which is freely available at http://159.226.118.31/PPI/index.html, promises to become useful in such areas as bio-analysis and drug development.

Keywords: amino acid distance frequency; amino acid index distribution; protein–protein interaction; pairwise kernel function; support vector machine

1. Introduction

Protein–protein interactions (PPIs) play an important role in such biological processes as host immune response, the regulation of enzymes, signal transduction and mediating cell adhesion. Understanding PPIs will bring more insight to disease etiology at the molecular level and potentially simplify the discovery of novel drug targets [1]. Information about protein–protein interactions have also been used to address many biological important problems [2–5], such as prediction of protein function [2], regulatory pathways [3], signal propagation during colorectal cancer progression [4], and identification of colorectal cancer related genes [5]. Experimental methods of identifying PPIs can be roughly categorized into low- and high-throughput methods [6]. However, PPI data obtained from low-throughput methods only cover a small fraction of the complete PPI network, and high-throughput methods often produce a high frequency of false PPI information [7]. Moreover, experimental methods are expensive, time-consuming and labor-intensive. The development of reliable computational methods to facilitate the identification of PPIs could overcome these obstacles.

Thus far, a number of computational approaches have been developed for the large-scale prediction of PPIs based on protein sequence, structure and evolutionary relationship in complete genomes. These methods can be roughly categorized into those that are genomic-based [8,9], structure-based [10], and sequence-based [11–26]. Genomic- and structure-based methods cannot be implemented if prior information about the proteins is not available. Sequence-based methods are more universal, but they concatenate the two feature vectors of protein P_a and P_b to represent the protein pair P_a–P_b, and the concatenation order of two feature vectors will affect the prediction results. For example, if we use feature vectors x_a, x_b to represent protein P_a and P_b, respectively, then the P_a–P_b protein pair can be expressed as $x_{ab} = x_a \oplus x_b$, or $x_{ba} = x_b \oplus x_a$. In general, however, $x_a \oplus x_b$ is not equal to $x_b \oplus x_a$. Furthermore, PPIs have a symmetrical character; that is, the interaction of protein P_a with protein P_b equals the interaction of protein P_b with protein P_a. Under these circumstances, concatenating two feature vectors of protein P_a and P_b to represent the protein pair P_a–P_b and then using the traditional kernel $k(x_1, x_2)$ to predict PPIs would not be workable.

Therefore, in this paper, we introduced two kinds of feature extraction approaches, amino acid distance frequency with PCA reducing the dimension (DFPCA) and amino acid index distribution (AAID) to represent the protein sequences, followed by the use of pairwise kernel function and SVM to predict PPI.

2. Results and Discussion

LIBSVM [27], loaded from http://www.csie.ntu.edu.tw/~cjlin, is a library for Support Vector Machines (SVMs), and it was used to design the classifier in this paper. The kernel program of the software was modified to the pairwise kernel functions, which were formed by the RBF genomic kernel function $K(x_1, x_2)$ in all experiments.

2.1. The Results of DFPCA and AADI with K_{II} Pairwise Kernel Function SVM

In statistical prediction, the following three cross-validation methods are often used to examine a predictor for its effectiveness in practical application: independent dataset test, *K*-fold crossover or subsampling test, and jackknife test [28]. However, of the three test methods, the jackknife test is deemed the least arbitrary that can always yield a unique result for a given benchmark dataset as demonstrated by Equations (28)–(30) in [29]. Accordingly, the jackknife test has been increasingly and widely used by investigators to examine the quality of various predictors (see, e.g., [30–41]). However, to reduce the computational time, we adopted the 10-fold cross-validation (10 CV) test in this study as done by many investigators with SVM as the prediction engine.

The four feature vector sets, Hf, Vf, Pf, and Zf, extracted with DFPCA and the five feature vector sets, LEWP710101, QIAN880138, NADH010104, NAGK730103 and AURR980116, extracted with AAID were employed as the input feature vectors for K_{II} pairwise radial basis kernel function (PRBF) SVM. The results of DFPCA and AAID are summarized in Table 1.

Table 1. Results of DFPCA and AAID with PRBF SVM in 10 CV test.

Feature Set	S_n (%)	*PPV* (%)	*ACC* (%)	*MCC*
Hf	95.94 ± 1.92	91.98 ± 2.88	93.78 ± 1.44	0.8765
Vf	95.66 ± 2.75	92.52 ± 2.40	93.96 ± 1.86	0.8798
Pf	95.78 ± 2.23	92.07 ± 1.69	93.76 ± 1.93	0.8760
Zf	96.06 ± 1.24	91.71 ± 3.13	93.69 ± 1.86	0.8747
LEWP710101	95.86 ± 2.23	92.08 ± 4.32	93.80 ± 2.42	0.8768
QIAN880138	96.06 ± 2.83	92.27 ± 1.50	94.00 ± 1.22	0.8808
NADH010104	95.82 ± 2.98	92.04 ± 2.51	93.76 ± 1.66	0.8760
NAGK730103	96.06 ± 2.83	92.09 ± 4.02	93.90 ± 3.31	0.8789
AURR980116	95.94 ± 2.07	92.33 ± 1.42	93.98 ± 1.24	0.8804

From Table 1, we can see that the performances of the two feature extraction approaches, *i.e.*, amino acid distance frequency with PCA (DFPCA) and amino acid index distribution (AAID), are nearly equal when using the K_{II} pairwise kernel SVM. The total prediction accuracies are 93.69%~94%. As previously noted, we used just five amino acid indices, including LEWP710101, QIAN880138, NADH010104, NAGK730103 and AURR980116, to produce the feature vector sets. When we tested the performance of AAID against the remaining 480 amino acid indices from AAindex, we found that the amino acid index does affect predictive results and that the total prediction accuracies of those amino acid indices were 79.4%~94%. Among our original five indices, as noted above, the performance of AAID was superior in comparison to the results from AAindex. To account for the better performance of our five indices, we point to the physicochemical and biochemical properties of amino acids. By single-linkage clustering, one of agglomerative hierarchical clustering methods, Tomii and Kanehisa [42] divided the minimum spanning of these amino acid indices into six regions: α and turn propensities, β propensity, amino acid composition, hydrophobicity, physicochemical properties, and other properties. The indices of LEWP710101, QIAN880138, NAGK730103 and AURR980116 are arranged into the region of α and turn propensities, while NADH010104 is arranged into the hydrophobicity region, indicating that the properties of α and turn propensities, and hydrophobicity contain more distinguishable information for predicting PPIs.

2.2. The Comparison of Pairwise Kernel Function with Traditional Kernel Function

In order to evaluate the performance of pairwise kernel function, we compared the results of pairwise radial basis kernel function (PRBF) and radial basis function kernel (RBF) with the same feature vector sets. For RBF, we concatenate the two feature vectors of protein P_a and protein P_b to represent the protein pair $P_a - P_b$; that is, feature vector $x_{ab} = x_a \oplus x_b$ was used as the input feature vector of RBF. The results of RBF and PRBF with DFPCA in the 10CV test are listed in Table 2.

Table 2. Results of RBF and PRBF with DFPCA in the 10 CV test.

Feature Set	Kernel Function	S_n (%)	*PPV* (%)	*ACC* (%)
Hf	RBF	89.96 ± 0.52	89.65 ± 2.17	89.88 ± 1.05
	PRBF	95.94 ± 1.92	91.98 ± 2.88	93.78 ± 1.44
Vf	RBF	90.20 ± 1.31	89.33 ± 2.60	89.72 ± 1.72
	PRBF	95.66 ± 2.75	92.52 ± 2.40	93.96 ± 1.86
Pf	RBF	89.32 ± 0.86	89.26 ± 2.91	89.28 ± 1.44
	PRBF	95.78 ± 2.23	92.07 ± 1.69	93.76 ± 1.93
Zf	RBF	90.84 ± 1.85	88.79 ± 2.50	89.64 ± 1.18
	PRBF	96.06 ± 1.24	91.71 ± 3.13	93.69 ± 1.86

Table 2 shows that the performance of PRBF is superior to that of RBF for predicting PPI. The total prediction accuracies of PRBF are higher at 3.9%~4.48% than those of RBF.

2.3. The Comparison of DF and DFPCA Feature Extraction Approaches

For the feature extraction approach of distance frequency of amino acids grouped with their physicochemical properties, we compared the results of DF and DFPCA with PRBF SVM to test the validity of adopting PCA. The reduced feature matrix is set to retain 99.9% information of the original feature matrix by PCA. The results of DF and DFPCA with PRBF SVM in the 10CV test are listed in Table 3.

Table 3. Results of DF and DFPCA with PRBF SVM in the 10 CV test.

Feature Set	Feature Extraction Approach	S_n (%)	*PPV* (%)	*ACC* (%)	*MCC*
Hf	DF	97.37 ± 2.55	66.67 ± 27.8	74.34 ± 24.3	0.5485
	DFPCA	95.94 ± 1.92	91.98 ± 2.88	93.78 ± 1.44	0.8765
Vf	DF	97.21 ± 2.39	71.40 ± 23.0	78.17 ± 27.1	0.6093
	DFPCA	95.66 ± 2.75	92.52 ± 2.40	93.96 ± 1.86	0.8798
Pf	DF	97.13 ± 4.70	69.48 ± 25.5	77.23 ± 27.2	0.5937
	DFPCA	95.78 ± 2.23	92.07 ± 1.69	93.76 ± 1.93	0.8760
Zf	DF	97.65 ± 4.82	62.29 ± 29.5	69.26 ± 23.6	0.4680
	DFPCA	96.06 ± 1.24	91.71 ± 3.13	93.69 ± 1.86	0.8747

From Table 3, we can see that the performance of DFPCA is superior to that of DF. The total prediction accuracies and *MCC* (see Equation (16) below) of DFPCA are 15.79%~24.43% and

0.2705~0.4067 higher than those of DF, respectively. Although the sensitivities of DF are a little higher (1.43%~1.59%) than those of DFPCA for the Hf, Vf, Pf and Zf feature sets, the positive predictive values are much less than that of DFPCA (21%~29%), which means that the DFPCA approach can largely reduce the false positives. These results show that the performance of DFPCA is superior to that of DF for predicting PPI. It should be noted that feature vectors generated with either DF or DFPCA contain statistical information of amino acids in protein sequences, as well as information about amino acid position and physicochemical properties.

2.4. The Performance of the Predictive System Influenced by Randomly Sampling the Noninteracting Protein Subchain Pairs

To investigate the influence of randomly sampling the noninteracting protein subchain pairs, we randomly sampled 2510 noninteracting protein subchain pairs five times to construct five negative sets, and we used the DFPCA approach with hydrophobicity property to predict PPI in the 10CV test. The results, as shown in Table 4, indicate that random sampling of the noninteracting protein subchain pairs in order to construct negative sets has little influence on the performance of the PPI-PKSVM.

Table 4. Effect of random sampling of the noninteracting protein subchain pairs on the performance of PPI-PKSVM with DFPCA and PRBF SVM in the 10CV test.

Sampling Time	S_n (%)	*PPV* (%)	*AAC* (%)	*MCC*
1	95.38 ± 3.35	91.20 ± 3.37	93.09 ± 3.45	0.8627
2	95.42 ± 1.39	91.52 ± 3.24	93.29 ± 1.65	0.8665
3	95.46 ± 3.03	91.21 ± 1.63	93.13 ± 2.29	0.8635
4	95.46 ± 3.03	91.49 ± 1.70	93.29 ± 2.13	0.8666
5	95.94 ± 1.92	91.98 ± 2.88	93.78 ± 1.44	0.8765

2.5. Comparison of Different Prediction Methods

To demonstrate the prediction performance of our method, we compared it with other methods [25] on a nonredundant dataset constructed by Pan and Shen [25], in which no protein pair has sequence identity higher than 25%. The number of positive links, *i.e.*, interacting protein pairs, is 3899, which is composed of 2502 proteins, and the number of negative links, *i.e.*, noninteracting protein pairs, is 4262, which is composed of 661 proteins. Among the prediction results of different methods shown in Table 5, the performance of PPI-PKSVM stands out as the best. When compared to Shen's LDA-RF, the accuracy (see Equation (15) below) and *MCC* of LEWP710101/QIAN880138 and Hf-DFPCA are respectively 1.9%, 2%, 0.038 and 0.039 higher. These results indicate that our method is a very promising computational strategy for predicting protein–protein interaction based on the protein sequences.

Table 5. Performance comparison of different PPI methods using Shen's dataset [a] in the 10 CV test.

Method	S_n (%)	S_p (%)	*ACC* (%)	*MCC*
LEWP710101	97.3 ± 0.04	99.2 ± 0.04	98.3 ± 0.00	0.966 ± 0.0006
QIAN880138	97.3 ± 0.10	99.1 ± 0.10	98.3 ± 0.10	0.966 ± 0.002
NADH010104	97.2 ± 0.07	99.2 ± 0.04	98.3 ± 0.05	0.965 ± 0.0007
NAGK730103	97.2 ± 0.06	99.2 ± 0.04	98.2 ± 0.06	0.965 ± 0.0004
AURR980116	97.3 ± 0.04	99.1 ± 0.06	98.2 ± 0.06	0.965 ± 0.0006
Hf-DFPCA	97.6 ± 0.20	99.1 ± 0.10	98.4 ± 0.10	0.967 ± 0.002
Vf-DFPCA	97.5 ± 0.10	98.9 ± 1.00	98.3 ± 0.80	0.965 ± 0.007
Pf-DFPCA	96.9 ± 0.10	99.5 ± 0.60	98.2 ± 0.60	0.964 ± 0.004
Zf-DFPCA	97.9 ± 0.90	96.0 ± 0.20	96.9 ± 1.10	0.939 ± 0.002
LDA-RF [b]	94.2 ± 0.40	98.0 ± 0.30	96.4 ± 0.30	0.928 ± 0.006
LDA-RoF [b]	93.7± 0.50	97.6 ± 0.60	95.7 ± 0.40	0.918 ± 0.007
LDA-SVM [b]	89.7 ± 1.30	91.5 ± 1.10	90.7 ± 0.90	0.813 ± 0.018
AC-RF [b]	94.0 ± 0.60	96.6 ± 0.40	95.5 ± 0.30	0.914 ± 0.007
AC-RoF [b]	93.3 ± 0.70	97.1 ± 0.70	95.1 ± 0.60	0.910 ± 0.009
AC-SVM [b]	94.0 ± 0.60	84.9 ± 1.70	89.3 ± 0.80	0.792 ± 0.014
PseAAC-RF [b]	94.1 ± 0.90	96.9 ± 0.30	95.6 ± 0.40	0.912 ± 0.007
PseAAC-RoF [b]	93.6 ± 0.90	96.7 ± 0.40	95.3 ± 0.50	0.907 ± 0.009
PseAAC-SVM [b]	89.9 ± 0.70	92.0 ± 0.40	91.2 ± 0.4	0.821 ± 0.006

[a] Shen's dataset contains two subdatasets, C and D, which are available at http://www.csbio.sjtu.edu.cn/bioinf/LR_PPI/Data.htm; [b] These results are taken from Table 4 of the literature [25].

3. Experimental Section

3.1. Dataset

To construct the PPI dataset, we first obtained the subchain pair name of PPIs from the PRISM (Protein Interactions by Structural Matching) server (http://prism.ccbb.ku.edu.tr/prism/), which was used to explore protein interfaces, and we downloaded the corresponding sequences of these protein subchain pairs from the Protein Data Bank (PDB) database (http://www.rcsb.org/pdb/). According to PRISM [43], a subchain pair is defined as an interacting subchain pair if the interface residues of two protein subchains exceed 10; otherwise, the subchain pair is defined as a noninteracting subchain pair. For example, suppose a protein complex has A, B, C and D subchains. If the interface residues of AB, AC, and BD subchain pairs total more than 10, while the interface residues of AD, BC and CD subchain pairs total less than 10, then the AB, AC, and BD subchain pairs are treated as interacting subchain pairs, while the AD, BC and CD subchain pairs are treated as noninteracting subchain pairs. All interacting protein subchain pairs were used in preparing the positive dataset, and all noninteracting subchain pairs were used in preparing the negative dataset. To reduce the redundancy and homology bias for methodology development, all protein subchain pairs were screened according to the following procedures [15]. (i) Protein subchain pairs containing a protein subchain with fewer than 50 amino acids were removed; (ii) For subchain pairs having ≥40% sequence identity, only one subchain pair was kept. The ≥40% determinant may be understood as follows. Suppose protein subchain pair A is formed with protein subchains A1 and A2 and protein subchain pair B is formed with protein subchains B1 and B2. If sequence identity between protein subchains A1 and B1 and A2 and B2 is ≥40%, or sequence identity between protein subchains A1 and B2 and between A2 and B1 is ≥40%, then the two protein subchain pairs are defined as having ≥40% sequence identity. In our method, we would only retain those subchain pairs having <40% sequence identity. After these screening procedures, the resultant positive set was comprised of 2510 interacting protein subchain pairs, while the resultant negative set contained many noninteracting protein subchain pairs. To avoid unbalanced data between the positive and negative sets, we randomly sampled the 2510 noninteracting protein subchain pairs to construct the negative set. Finally, a PPI dataset consisting of 2510 PPI subchain pairs and 2510 noninteracting protein subchain pairs was constructed.

3.2. Distance Frequency of Amino Acids Grouped with Their Physicochemical Properties

The frequency of the distance between two successive amino acids, or distance frequency, was used to predict subcellular location by Matsuda *et al.*, [44] and can be described as follows: For a protein sequence *P*, the distance set d_A between two successive letters (e.g., *A*) appearing in protein sequence *P* can be represented as:

$$d_A = \{d_1, d_2, ..., d_i, ..., d_{n_A-1}\} \quad i = 1, ...n_A - 1 \tag{1}$$

where n_A is number of letter *As* appearing in protein sequence *P*, d_i is the distance from the *i*th letter *A* to the $(i + 1)$th letter *A*, and d_i is calculated in a left-to-right fashion. The distance frequency vector for letter *A* can be defined by the following equation:

$$f_A = [N_1, N_2, \cdots, N_j, \cdots N_m] \tag{2}$$

where N_j represents the number of times that the *j*th distance unit appears in the d_A set. For example, considering the protein sequence *AACDAMMADA*, the distance sets of letters *A*, *C*, *D* and *M* are shown respectively as

$$d_A = \{1,3,3,2\}, d_C = \{0\}, d_D = \{5\}, d_M = \{1\}$$

As a result, the corresponding distance frequency vectors are shown respectively as $Df_A = [1,1,2,0,0], Df_C = [0,0,0,0,0], Df_D = [0,0,0,0,1], Df_M = [1,0,0,0,0]$. The other 16 basic amino acid distance frequency vectors are zero vector, or $V = [0,0,0,0,0]$. Thus, we can use the feature vector x to encode the protein sequence *P*:

$$x = [Df_A, Df_C, Df_D, \cdots, Df_Y]$$

In this work, we used the concept of distance frequency [44] and borrowed Dubchak's idea of representing the amino acid sequence with four physicochemical properties [45] to encode the protein subchain sequence. First, according to the amino acid value given by such physicochemical properties as hydrophobicity [46], normalized van der Waals volume [47], polarity [48] and polarizability [49], the 20 natural amino acids can be divided into three groups [45], as listed in the Table 6. For Hydrophobicity, Normalized van der Waals Volume, Polarity and Polarizability, the amino acids in Group 1, Group 2 and Group 3 were expressed as H_1, H_2, H_3; V_1, V_2, V_3; P_1, P_2, P_3; and Z_1, Z_2 and Z_3, respectively. Second, each protein subchain sequence was then translated into the appropriate three-symbol sequence, depending on the particular physicochemical property, be it $H_{1\text{–}3}$, $V_{1\text{–}3}$, $P_{1\text{–}3}$, or $Z_{1\text{–}3}$. For example, suppose that the original protein sequence is MKEKEFQSKP. Then, by the set of symbols denoted above, in this case, hydrophobicity, this sequence can be translated into $H_3H_1H_1H_1H_1H_3H_1H_2H_1H_2$, and the same would be true for $V_{1\text{–}3}$, $P_{1\text{–}3}$, or $Z_{1\text{–}3}$. Third, the distance frequency of every symbol in the translated sequence was computed. In the above example, the H_1, H_2, H_3 distance frequency would be respectively computed for the sequence $H_3H_1H_1H_1H_1H_3H_1H_2H_1H_2$. Finally, every protein subchain sequence can be encoded by the following feature vector:

$$x_H = [x_{H_1}, x_{H_2}, x_{H_3}]^T, x_V = [x_{V_1}, x_{V_2}, x_{V_3}]^T, x_P = [x_{P_1}, x_{P_2}, x_{P_3}]^T, x_Z = [x_{Z_1}, x_{Z_2}, x_{Z_3}]^T \quad (3)$$

Table 6. Amino acid groups classified according to their physicochemical value.

Physicochemical property	**Group 1**	**Group 2**	**Group 3**
Hydrophobicity	H_1: R,K,E,D,Q,N	H_2: G,A,S,T,P,H,Y	H_3: C,V,L,I,M,F,W
van der Waals volume	V_1: G,A,S,C,T,P,D	V_2: N,V,E,Q,I,L	V_3: M,H,K,F,R,Y,W
Polarity	P_1: L,I,F,W,C,M,V,Y	P_2: P,A,T,G,S	P_3: H,Q,R,K,N,E,D
Polarizability	Z_1: G,A,S,D,T	Z_2: C,P,N,V,E,Q,I,L	Z_3: K,M,H,F,R,Y,W

Conveniently, the feature set based on hydrophobicity, normalized van der Waals volume, polarity, and polarizability can be written as Hf, Vf, Pf and Zf, respectively. In general, the dimensions of two feature vectors generated separately by two protein subchains are unequal. To solve this issue, we enlarge the feature vector dimension of one protein subchain such that it has a feature vector dimension equal to that of another subchain. For example, given the following protein subchain pair $P_a - P_b$:

Subchain P_a amino acid sequence: MKEKEFQSKP
Subchain P_b amino acid sequence: QNSLALHKVIMVGSG

If we adopt the property of hydrophobicity, then P_a and P_b amino acid sequences can be translated into the following symbol sequence, respectively.

Subchain P_a: $H_3H_1H_1H_1H_1H_3H_1H_2H_1H_2$
Subchain P_b: $H_1H_1H_2H_3H_2H_3H_2H_1H_3H_3H_3H_3H_2H_2H_2$

Then, the distance sets of subchains P_a and P_b are shown as: $d^a_{H_1} = \{1,1,1,2,2\}, d^a_{H_2} = \{2\}, d^a_{H_3} = \{5\}, d^b_{H_1} = \{1,6\}, d^b_{H_2} = \{2,2,6,1,1\}, d^b_{H_3} = \{2,3,1,1,1,\}$, and the distance frequency vectors of subchains P_a and P_b are as follows:

$$x_a = [x^a_{H_1}, x^a_{H_2}, x^a_{H_3}], x_b = [x^b_{H_1}, x^b_{H_2}, x^b_{H_3}]$$

where

$$x^a_{H_1} = [3,2,0,0,0,0], x^a_{H_2} = [0,1,0,0,0,0], x^a_{H_3} = [0,0,0,0,1,0],$$
$$x^b_{H_1} = [1,0,0,0,0,1], x^b_{H_2} = [2,2,0,0,0,1], \; x^b_{H_3} = [3,1,1,0,0,0]$$

Hereinafter we will use “DF” to represent the distance frequency method by grouping amino acids with their physicochemical properties.

By our use of DF to represent the protein subchain pair, we can see that the feature vector is sparse, while the vector dimension is large, when the subchain sequence is longer. To further extract the features, Principal Component Analysis (PCA) was then used to reduce the dimension, and amino acid distance frequency combined with PCA reducing the dimension is now termed DFPCA.

3.3. Amino Acid Index Distribution (AAID)

Let $I_1, I_2, \ldots, I_i, \cdots, I_{20}$ be the amino acid physicochemical value of the 20 natural amino acids α_i (A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, and Y), respectively, which can be accessed through the DBGET/LinkDB system by inputting an amino acid index (e.g., LEWP710101). An amino acid index is a set of 20 numerical values representing any of the different physicochemical and biochemical properties of amino acids. We can download these indices from the AAindex database (http://www.genome.jp/aaindex/).

For a given protein sequence *P* whose length is L, we replace each residue in the primary sequence by its amino acid physicochemical value, which results in a numerical sequence $h_1, h_2, \ldots, h_l, \ldots, h_L$, $(h_l \in I_1, I_2, \ldots, I_{20})$.

Then, we can define the following feature w_i of amino acid α_i to represent the protein sequences:

$$w_i = I_i \bullet f_i \tag{4}$$

Where f_i is the frequency of amino acid α_i that occurs in protein sequecne *P*, I_i is the physicochemical value of amino acid α_i, and the symbol • indicates the simple product. f_i and I_i are mutually independent. Obviously, w_i includes the physicochemical information and statistical information of amino acid α_i, but it loses the sequence-order information. Therefore, to let feature vectors contain more sequence-order information, we introduced the 2-order center distance d_i by considering the position of amino acid α_i, which is defined as

$$d_i = \sum_{j=1}^{N_{\alpha_i}} \left(\frac{k_{i,j} - \bar{k}_i}{L} \bullet I_i \right)^2 \tag{5}$$

where N_{α_i} is the total number of amino acid α_i appearing in the protein sequence *P*, $k_{i,j}$ $(j = 1, 2, \cdots, N_{\alpha_i})$ is the jth position of the amino acid α_i in the sequence, and $\bar{k}_i$ is the mean of the position of amino acid α_i.

Now feature d_i contains the physicochemical information, statistical information and the sequence-order information of amino acid α_i, but it still does not distinguish the protein pairs in some cases. For example, assume two protein pairs $P_a - P_b$ and $P_c - P_d$. The sequences of protein P_a, P_b, P_c and P_d are respectively shown as:

P_a: MPPRNKPNRR; P_b: MPNPRNNKPPGRKTR
P_c: MPRRNPPNRK; P_d: MGTRPPRNNKPNPRK

Obviously, P_a and P_c, as well as P_b and P_d, have the same w_i and d_i. If we use the orthogonal sum vector, we cannot distinguish between the $P_a - P_b$ and $P_c - P_d$ protein pairs. To solve this problem, the 3-order center distance t_i of amino acid α_i was introduced, which is defined as

$$t_i = \sum_{j=1}^{N_{\alpha_i}} \left(\frac{k_{i,j} - \bar{k}_i}{L} \bullet I_i \right)^3 \tag{6}$$

Finally, we can use a combined feature vector to represent protein sequence *P* by serializing above three features as

$$x = [w_1, \cdots, w_i, \cdots, w_{20}, d_1, \cdots, d_i, \ldots, d_{20}, t_1, \cdots, t_i, \cdots, t_{20}]^T \tag{7}$$

The protein pair $P_a - P_b$ can now be represented by the following feature vectors:

$$x_{ab} = [w_1^a, \cdots, w_{20}^a, d_1^a, \cdots d_{20}^a, t_1^a, \cdots, t_{20}^a, w_1^b, \cdots, w_{20}^b, d_1^b, \cdots, d_{20}^b, t_1^b, \cdots, t_{20}^b]^T \tag{8}$$

or

$$x_{ba} = [w_1^b, \cdots, w_{20}^b, d_1^b, \cdots d_{20}^b, t_1^b, \cdots, t_{20}^b, w_1^a, \cdots, w_{20}^a, d_1^a, \cdots, d_{20}^a, t_1^a, \cdots, t_{20}^a]^T \tag{9}$$

Generally, vector x_{ab} is not equal to vector x_{ba}. As such, if a query protein pair $P_a - P_b$ is represented by x_{ab} and x_{ba} respectively, the prediction results may be different. In this paper, we will choose the pairwise kernel function to solve this dilemma.

3.4. Pairwise Kernel Function

Ben-Hur and Noble [13] first introduced a tensor product pairwise kernel function K_{I} to measure the similarity between two protein pairs. The comparison between a pair (x_1, x_2) and another pair (x_3, x_4) for K_{I} is done through the comparison of x_1 with x_3 and x_2 with x_4, on the one hand, and the comparison of x_1 with x_4 and x_2 with x_3, on the other hand, as

$$K_{\mathrm{I}}((x_1, x_2), (x_3, x_4)) = K(x_1, x_3) \cdot K(x_2, x_4) + K(x_1, x_4) \cdot K(x_2, x_3) \tag{10}$$

However, the K_{I} kernel does not consider differences between the elements of comparison pairs in the feature space; therefore, Vert [50] proposed the following metric learning pairwise kernel K_{II}:

$$K_{\mathrm{II}}((x_1, x_2), (x_3, x_4)) = (K(x_1, x_3) + K(x_2, x_4) - K(x_1, x_4) - K(x_2, x_3))^2 \tag{11}$$

In particular, two protein pairs might be very similar for the K_{II} kernel, even if the patterns of the first protein pair are very different from those of the second protein pair, whereas the K_{I} kernel could result in a large dissimilarity between the two protein pairs. It is easy to prove that the K_{II} kernel satisfies both Mercer's condition and the pairwise kernel function condition. In this paper, we use the K_{II} kernel function to predict PPI.

3.5. Assessment of Prediction System

Sensitivity (S_n), specificity (S_p), positive predictive value (*PPV*) and total prediction accuracy (*ACC*) [39–41] were employed to measure the performance of PPI-PKSVM.

$$S_n = \frac{TP}{TP + FN} \tag{12}$$

$$S_p = \frac{TN}{TN + FP} \tag{13}$$

$$PPV = \frac{TP}{TP + FP} \tag{14}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{15}$$

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

where *TP* and *TN* are the number of correctly predicted subchain pairs of interacting proteins and noninteracting proteins, respectively, and *FP* and *FN* are the number of incorrectly predicted subchain pairs of noninteracting proteins and interacting proteins, respectively.

4. Conclusions

In this work, we introduced two feature extraction approaches to represent the protein sequence. One is amino acid distance frequency with PCA reducing the dimension, termed DFPCA. Another is amino acid index distribution based on the physicochemical values of amino acids, termed AAID. The pairwise kernel function SVM was employed as the classifier to predict the PPIs. From the results, we can conclude that (i) the performance of DFPCA is better than that of DF; (ii) the prediction power of PRBF is superior to RBF, suggesting that designing a rational pairwise kernel function is important for predicting PPIs; (iii) DFPCA and AAID with pairwise kernel function SVM are effective and promising approaches for predicting PPIs and may complement existing methods. Since user-friendly and publicly accessible web servers represent the future direction in the development of predictors, we have provided a web server for PPI-PKSVM, and it can be found at (http://159.226.118.31/PPI/index.html). PPI-PKSVM in its present version can be used to evaluate one protein pair. However, we will soon be developing a newer online version able to predict large numbers of PPIs.

Acknowledgments

This paper was supported by the National Natural Science Foundation of China (No. 61170134 and 60775012).

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Lucy, S.; Harpreet, K.S.; Gary, D.B.; Anton, J.E. Computational prediction of protein–protein interactions. *Mol. Biotechnol.* **2008**, *38*, 1–17.
2. Hu, L.; Huang, T.; Shi, X.; Lu, W.C.; Cai, Y.D.; Chou, K.C. Predicting functions of proteins in mouse based on weighted protein–protein interaction network and protein hybrid properties. *PLoS One* **2011**, *6*, e14556.
3. Huang, T.; Chen, L.; Cai, Y.D.; Chou, K.C. Classification and analysis of regulatory pathways using graph property, biochemical and physicochemical property, and functional property. *PLoS One* **2011**, *6*, e25297.

4. Jiang, Y.; Huang, T.; Chen, L.; Gao, Y.F.; Cai, Y.D.; Chou, K.C. Signal propagation in protein interaction network during colorectal cancer progression. *BioMed Res. Int.* **2013**, *2013*, doi:10.1155/2013/287019.
5. Li, B.Q.; Huang, T.; Cai, Y.D.; Chou, K.C. Identification of colorectal cancer related genes with mRMR and shortest path in protein–protein interaction network. *PLoS One* **2013**, *7*, e33393.
6. Shoemaker, B.A.; Panchenko, A.R. Deciphering protein–protein interactions. Part I Experimental techniques and databases. *PLoS Comput. Biol.* **2007**, *3*, e42.
7. Han, J.D.; Dupuy, D.; Bertin, N.; Cusick, M.E.; Vidal, M. Effect of sampling on topology predictions of protein–protein interaction networks. *Nat. Biotechnol.* **2005**, *23*, 839–844.
8. Marcotte, E.M.; Pellegrini, M.; Ng, H.L.; Rice, D.W.; Yeates, T.O.; Eisenberg, D. Detecting protein function and protein–protein interactions from genome sequences. *Science* **1999**, *285*, 751–753.
9. Juan, D.; Pazos, F.; Valencia, A. High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. *Proc. Natl. Acad. Sci.USA* **2008**, *105*, 934–939.
10. Singhal, M.; Resat, H. A domain-based approach to predict proteinprotein interactions. *BMC Bioinforma.* **2007**, *8*, 199.
11. Bock, J.R.; Gough, D.A. Predicting protein–protein interactions from primary structure. *Bioinformatics* **2001**, *17*, 455–460.
12. Gomez, S.M.; Noble, A.S.; Rzhetsky, A. Learning to predict protein–protein interactions from protein sequences. *Bioinformatics* **2003**, *19*, 1875–1881.
13. Ben-Hur, A.; Noble, W.S. Kernel methods for predicting protein–protein interactions. *Bioinformatics* **2005**, *21*, i38–i46.
14. Martin, S.; Roe, D.; Faulon, J.L. Predicting protein–protein interactions using signature products. *Bioinformatics* **2005**, *21*, 218–226.
15. Chou, K.C.; Cai, Y.D. Predicting protein–protein interactions from sequences in a hybridization space. *J. Proteome Res*. **2006**, *5*, 316–322.
16. Nanni, L.; Lumini, A. An ensemble of K-local hyperplanes for predicting protein–protein interactions. *Bioinformatics* **2006**, *22*, 1207–1210.
17. Pitre, S.; Dehne, F.; Chan, A.; Cheetham, J.; Duong, A.; Emili, A.; Gebbia, M.; Greenblatt, J.; Jessulat, M.; Krogan, N.; *et al*. PIPE: A protein–protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs. *BMC Bioinforma.* **2006**, *7*, 365.
18. Li, X.L.; Tan, S.H.; Ng, S.K. Improving domain-based protein interaction prediction using biologically-significant negative dataset. *Int. J. Data Min. Bioinforma.* **2006**, *1*, 138–149.
19. Shen, J.W.; Zhang, J.; Luo, X.M.; Zhu, W.L.; Yu, K.Q.; Chen, K.X.; Li, Y.X.; Jiang, H.L. Predicting protein–protein interactions based only on sequences information. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 4337–4341.
20. Guo, Y.Z.; Yu, L.Z.; Wen, Z.N.; Li, M.L. Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences. *Nucleic Acids Res*. **2008**, *36*, 3025–3030.
21. Chen, X.W.; Han, B.; Fang, J.; Haasl, R.J. Large-scale protein–protein interaction prediction using novel kernel methods. *Int. J. Data Min. Bioinforma.* **2008**, *2*, 145–156.

22. Chen, W.; Zhang, S.W.; Cheng, Y.M.; Pan, Q. Prediction of protein–protein interaction types using the decision templates based on multiple classier fusion. *Math. Comput. Model.* **2010**, *52*, 2075–2084.
23. Guo, Y.; Li, M.; Pu, X.; Li, G.; Guang, X.; Xiong, W.; Li, J. PRED_PPI: A server for predicting protein–protein interactions based on sequence data with probability assignment. *BMC Res. Notes* **2010**, *3*, 145.
24. Yu, C.Y.; Chou, L.C.; Chang, D.T.H. Predicting protein–protein interactions in unbalanced data using the primary structure of proteins. *BMC Bioinforma.* **2010**, *11*, 167.
25. Pan, X.Y.; Zhang, Y.N.; Shen, H.B. Large-scale prediction of human protein–protein interactions from amino acid sequence based on latent topic features. *J. Proteome Res.* **2010**, *9*, 4992–5001.
26. Liu, C.H.; Li, K.C.; Yuan, S. Human protein–protein interaction prediction by a novel sequence-based co-evolution method: Co-evolutionary divergence. *Bioinformatics* **2013**, *29*, 92–98.
27. Hsu, C.; Lin, C.J. A comparision of methods for multi-class support vector machines. *IEEE Trans. Neural Netw.* **2002**, *3*, 415–425.
28. Chou, K.C.; Zhang, C.T. Review: Prediction of protein structural classes. *Crit. Rev. Biochem. Mol. Biol.* **1995**, *30*, 275–349.
29. Chou, K.C. Some remarks on protein attribute prediction and pseudo amino acid composition (50th Anniversary Year Review). *J. Theor. Biol.* **2011**, *273*, 236–247.
30. Esmaeili, M.; Mohabatkar, H.; Mohsenzadeh, S. Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses. *J. Theor. Biol.* **2010**, *263*, 203–209.
31. Hajisharifi, Z.; Piryaiee, M.; Mohammad Beigi, B.; Mandana, B.; Hassan, M. Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity via Ames test. *J. Theor. Biol.* **2014**, *341*, 34–40.
32. Mohabatkar, H.; Mohammad Beigi, M.; Esmaeili, A. Prediction of GABA(A) receptor proteins using the concept of Chou's pseudo-amino acid composition and support vector machine. *J. Theor. Biol.* **2011**, *281*, 18–23.
33. Xu, Y.; Ding, J.; Wu, L.Y.; Chou, K.C. iSNO-PseAAC: Predict cysteine *S*-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition. *PLoS One* **2013**, *8*, e55844.
34. Xu, Y.; Shao, X.J.; Wu, L.Y.; Deng, N.Y.; Chou, K.C. iSNO-AAPair: Incorporating amino acid pairwise coupling into PseAAC for predicting cysteine *S*-nitrosylation sites in proteins. *PeerJ* **2013**, *1*, e171.
35. Chen, W.; Feng, P.M.; Lin, H.; chou, K.C. iRSpot-PseDNC: Identify recombination spots with pseudo dinucleotide composition. *Nucleic Acids Res.* **2013**, *41*, e69.
36. Qiu, W.R.; Xiao, X.; Chou, K.C. iRSpot-TNCPseAAC: Identify recombination spots with trinucleotide composition and pseudo amino acid components. *Int. J. Mol. Sci.* **2014**, *15*, 1746–1766.
37. Min, J.L.; Xiao, X.; Chou, K.C. iEzy-Drug: A web server for identifying the interaction between enzymes and drugs in cellular networking. *Biomed. Res. Int*. **2013**, *2013*, 701317.

38. Zhang, S.W.; Liu, Y.F.; Yu, Y.; Zhang, T.H.; Fan, X.N. MSLoc-DT: A new method for predicting the protein subcellular location of multispecies based on decision templates. *Anal. Biochem.* **2014**, *449*, 164–171.
39. Chen, W.; Zhang, S.W.; Cheng, Y.M.; Pan, Q. Identification of protein-RNA interaction sites using the information of spatial adjacent residues. *Proteome Sci.* **2011**, *9*, S16.
40. Zhang, S.W.; Zhang, Y.L.; Yang, H.F.; Zhao, C.H.; Pan, Q. Using the concept of Chou's pseudo amino acid composition to predict protein subcellular localization: An approach by incorporating evolutionary information and von Neumann entropies. *Amino Acids* **2008**, *34*, 565–572.
41. Zhang, S.W.; Chen, W.; Yang, F.; Pan, Q. Using Chou's pseudo amino acid composition to predict protein quaternary structure: A sequence-segmented PseAAC approach. *Amino Acids* **2008**, *35*, 591–598.
42. Tomii, K.; Kanehisa, M. Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. *Protein Eng.* **1996**, *9*, 27–36.
43. Ogmen, U.; Keskin, O.; Aytuna, A.S.; Nussinov, R.; Gürsoy, A. PRISM: Protein interactions by structural matching. *Nucleic Acids Res.* **2005**, *33*, 331–336.
44. Matsuda, S.; Vert, J.P.; Saigo, H.; Ueda, N.; Toh, H.; Akutsu, T. A novel representation of protein sequences for prediction of subcellular location using support vector machines. *Protein Sci.* **2005**, *14*, 2804–2813.
45. Dubchak, I.; Muchnik, I.; Mayor, C.; Dralyuk, I.; Kim, S.H. Recognition of a protein fold in the context of the SCOP classification. *Proteins* **1999**, *35*, 401–407.
46. Chothia, C.; Finkelstein, A.V. The classification and origins of protein folding patterns. *Annu. Rev. Biochem*. **1999**, *59*, 1007–1039.
47. Fauchere, J.L.; Charton, M.; Kier, L.B.; Verloop, A.; Pliska, V. Amino acid side chain parameters for correlation studies in biology and pharmacology. *Int. J. Peptide Protein Res.* **1998**, *32*, 269–278.
48. Grantham, R. Amino acid difference formula to help explain protein evolution. *Science* **1974**, *185*, 862–864.
49. Charton, M.; Charton, B.I. The structural dependence of amino acid hydrophobicity parameters. *J. Theor. Biol*. **1982**, *99*, 629–644.
50. Vert, J.P.; Qiu, J.; Noble, W.S. A new pairwise kernel for biological network inference with support vector machines. *BMC Bioinforma.* **2007**, *8*, S8.