Prediction of Protein-Protein Interactions from Protein Sequence Using Local Descriptors

Lei Yang[1,2], Jun-Feng Xia[1,2,*] and Jie Gui[2,3]

[1]*Intelligent Computing Laboratory, Hefei Institute of Intelligent Machines, Chinese Academy of Sciences, P.O. Box 1130, Hefei, Anhui 230031, China;* [2]*School of Life Science, University of Science and Technology of China, Hefei, Anhui 230027,China;*[3]*Department of Automation, University of Science and Technology of China, Hefei, Anhui 230026, China*

Abstract: With a huge amount of protein sequence data, the computational method for protein–protein interaction (PPI) prediction using only the protein sequences information have drawn increasing interest. In this article, we propose a sequence-based method based on a novel representation of local protein sequence descriptors. Local descriptors account for the interactions between residues in both continuous and discontinuous regions of a protein sequence, so this method enables us to extract more PPI information from the sequence. A series of elaborate experiments are performed to optimize the prediction model by varying the parameter k and the distance measuring function of the k-nearest neighbors learning system and the ways of coding a protein pair. When performed on the PPI data of *Saccharomyces cerevisiae*, the method achieved 86.15% prediction accuracy with 81.03% sensitivity at the precision of 90.24%. An independent data set of 986 *Escherichia coli* PPIs was used to evaluate this prediction model and the prediction accuracy is 73.02%. Given the complex nature of PPIs, the performance of our method is promising, and it can be a helpful supplement for PPIs prediction.

Keywords: Feature representation, KNNs, local descriptors, PPIs prediction, protein sequence, sequence-based method.

INTRODUCTION

Identification of protein–protein interactions (PPIs) is crucial for elucidating protein functions and further understanding various biological processes in a cell. It has been the focus of the post-proteomic researches.

Recent years, high throughput technologies have been developed for the large-scale PPI analysis, such as yeast two-hybrid screening methods [1], immunoprecipitation [2], protein chips [3]. However, there are some disadvantages for existing experimental methods, such as time-intensive, high cost and a small fraction of the complete PPI network covered. In addition, these approaches suffer from high rates of both false negative and false positive predictions [1, 4]. So the computational approaches are a necessary complement for protein interaction discovery.

A number of computational methods have been developed for the prediction of PPIs based on various data types, including the gene fusion/Rosetta Stone method [5], the phylogenetic profile method [6, 7], the interacting proteins coevolution method [8-10], and the literature mining method [11]. There are also methods that integrate interaction information from many different data sources [12]. However, these methods are not universal, because the accuracy and reliability of these methods depend on the prior information of the protein pairs. In addition, compared to the huge amount of available protein sequences, others data capable for PPIs prediction are scarce. For example, by July 29, 2008, there are 392 667 identified protein sequences in Uniprot/Swissprot (reviewed, manually annotated) and only 47 978 known protein structures in PDB [13]. Therefore the sequence-based methods are much more universal and acceptable.

Recently, many methods have been proposed for inferring PPIs from protein sequence. These previous studies differ in protein feature presentation or in machine learning system, including using a support vector machine (SVM) with several structural and physiochemical descriptors [14, 15], using Bayesian network based on codon usage [16], using k-nearest neighbors (KNNs) learning system based on pseudo amino acid composition [17], using rotation forest based on autocorrelation descriptor [18] and so on. One of the excellent works is a SVM-based method proposed by Shen *et al*. [19]. In this method, the 20 amino acids were clustered into seven classes according to their dipoles and volumes of the side chains, and then the conjoint triad method abstracts the features of protein pairs based on the classification of amino acids. When applied to predicting human PPIs, this method yields a high prediction accuracy of 83.9%. However, Guo *et al*. [20] pointed out that the conjoint triad method can not takes neighbouring effect into account and the interactions usually occur in the discontinuous amino acids segments in the sequence. They proposed a method based on SVM and auto covariance (AC) to extract the interactions information in the discontinuous amino acids segments in the sequence. This approach yielded a prediction accuracy of 86.55%, when tested by the non-redundant data of *Saccharomyces cerevisiae*.

*Address correspondence to this author at the Intelligent Computing Lab, Hefei Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei, Anhui, 230031, China; Tel: +86-551-5591108; Fax: +86-551-5592751; E-mail: jfxia@mail.ustc.edu.cn

0929-8665/10 $55.00+.00

Though the above sequence-based methods are useful, one flaw of them is that the interactions information can't be drawn from both continuous and discontinuous amino acids segments at the same time. To overcome this problem, we propose a sequence-based method based on a novel representation of local protein sequence descriptors. Local descriptors have been used for the classification of several protein families [21, 22]. A comparative method in the context of protein family classification can be found in [23]. Local descriptors in such context have the following biological explanation: local descriptors as a method for predicting allergenic potential can address the dual issues of conformational and overlapping B-cell epitope recognition sites by identifying both continuous and discontinuous B-cell epitope binding patterns in allergenic proteins [22].

The effectiveness of local descriptors depends largely on the correct selection of amino acid grouping [21]. By grouping amino aids into a reduced alphabet, we can create a more accurate protein sequence representation. A number of groupings have been introduced based on biochemical properties of the amino acids [19, 21, 23, 24]. Here, we adopt the amino acids grouping according to the successful use of classification in [19]. We combine this classification with local descriptors [21, 22] for a novel representation of local protein sequence descriptors, which enabling us to extract more PPI information from the sequence with the purpose of improving the PPIs prediction.

In this work, we first represent each protein sequence as a vector by using the proposed novel representation of local protein sequence descriptors, and then characterize a protein pair in different feature vectors by coding the vectors of two proteins in this protein pair. Finally, a KNNs model is constructed using these feature vectors of the protein pair as input. When performed on the PPI data of *S. cerevisiae* [20], our method achieved 86.15% prediction accuracy with 81.03% sensitivity at the precision of 90.24%. The prediction model was further assessed using the independent data set of the *Escherichia coli* PPIs and yielded 73.02% prediction accuracy. The performance of the combination between the proposed representation of local protein sequence descriptors and SVM was also evaluated using the independent dataset from *E. coli* and an accuracy of 71.24% was obtained.

METHODS

Classification of Amino Acids

According to the method of Shen *et al.* [19], the 20 amino acids are firstly clustered into seven classes based on the dipoles and volumes of the side chains in order to reduce the dimensions of vector space and to suit synonymous mutation. The amino acid groups are shown in Table **1**. Then, every amino acid in each protein sequence is replaced by the index depending on its grouping. For example, protein sequence KLLSHCLLVTLAAHLPAEFTPAV is replaced by 52234722132114221623211 based on this classification of amino acids.

Local Descriptors

Local descriptors are an alignment-free approach used previously to classify several protein families [21, 22]. Coupled with an educated and underlying classification of amino acids, the techniques using local descriptors can create more accurate protein sequence representation.

Table 1. Classification of Amino Acids

No.	Class
1	Ala, Gly, Val
2	Ile, Leu, Phe, Pro
3	Tyr, Met, Thr, Ser
4	His, Asn, Gln, Tpr
5	Arg, Lys
6	Asp, Glu
7	Cys

Feature vector extraction using local descriptors in this work is similar to the one discussed in earlier studies [21, 22, 25].

Each protein is split into 10 local regions of varying length and composition in order to better capture PPI information from both continuous and discontinuous amino acids segments of the sequence Fig. (**1**). For each local region, three descriptors, *composition* (*C*), *transition* (*T*) and *distribution* (*D*), are calculated based on the definition given as follows.

Composition

Composition represents the proportion of each amino acid group.

Transition

Transition is represented by the percentage frequency with which amino acids in one group are followed by amino acids in another groups.

Distribution

Distribution descriptor measures proportion of the entire sequence within which the first, 25%, 50%, 75% and 100% of the amino acids in a particular group are located.

Protein Feature Representation and Coding Protein Pairs

In this article, we present a novel protein feature reresentation by combining local descriptors with the amino acids classification method discussed above, in which the 20 amino acids were clustered into seven groups according to the dipoles and volumes of the side chains. For each local region, the three descriptors (*C*, *T* and *D*) were calculated and concatenated, and a total of 63 descriptors are generated: 7 for *C*, 21 ((7×6)/2) for *T* and 35 (7×5) for *D*. Then, all descriptors from local regions were concatenated and a total 630-dimensional vector has been built to represent each protein sequence.

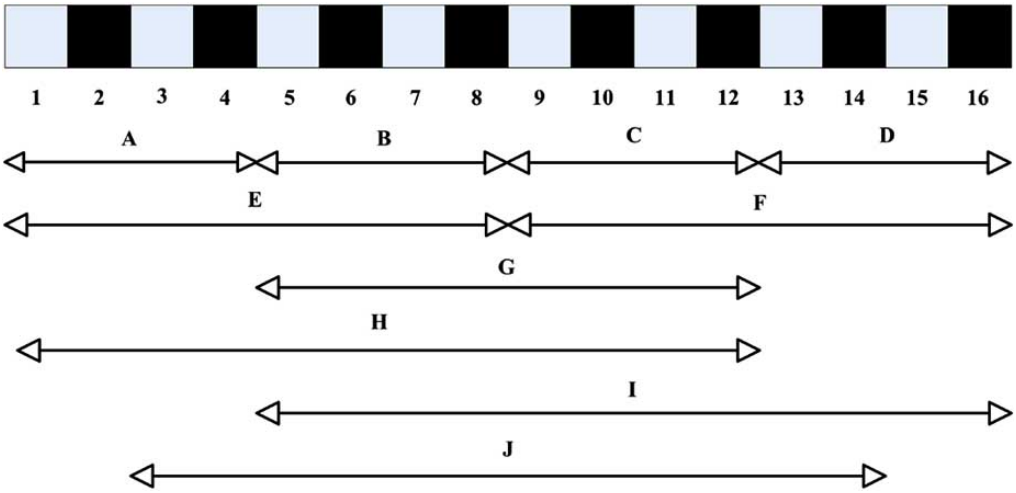

Figure 1. The 10 descriptor regions (**A**–**J**) for a theoretical protein sequence. The regions (**A**–**D**) and (**E**–**F**) are generated by dividing the entire sequence into four equal regions and into two equal regions respectively. **G** represents the central 50% of the sequence, **H** the first 75%, **I** the final 75% and **J** the central 75%.

Here, A and B are defined as the vector space of the protein *a* and protein *b* respectively. Thus, the size of either A or B is 630. In order to give more effectively and adequately presentation of the PPIs information, we code the protein-protein pair in four different means, which are defined as follows.

Protein *a*: $A = (x_1, x_2, \cdots, x_{630})$; Protein *b*: $B = (y_1, y_2, \cdots, y_{630})$.

Cod1: $\text{abs(A-B)} = \left[|x_1 - y_1|, |x_2 - y_2|, \cdots, |x_{630} - y_{630}|\right]$

Cod2: $\text{(A+B)} = \left[(x_1 + y_1), (x_2 + y_2), \cdots, (x_{630} + y_{630})\right]$

Cod3: $\left(\text{abs(A-B)} \oplus \text{(A+B)}\right) = \left[|x_1 - y_1|, |x_2 - y_2|, \cdots, |x_{630} - y_{630}|, (x_1 + y_1), (x_2 + y_2), \cdots, (x_{630} + y_{630})\right]$

Cod4: $(A \oplus B) = [x_1, x_2, \cdots, x_{630}, y_1, y_2, \cdots, y_{630}]$

The K-Nearest Neighbors (KNNs) Algorithm

The prediction was performed with the KNNs [26], which is a popular and widely applied algorithm to researchers. It is not theoretically described in this article. More details about KNNs can be found in many machine learning publications.

In order to promote the performance of KNNs, we adopt three kinds of distance measuring functions, which are defined as follows.

$x = (x_1, x_2, \cdots, x_n)', y = (y_1, y_2, \cdots, y_n)'.$

The *1-norm distance* (*L1*) : $d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$

The *2-norm distance* (*L2*) :

$$d(x, y) = \left[\sum_{i=1}^{n} (x_i - y_i)^2\right]^{1/2}$$

The cosine similarity (*Cos*) :

$$cox(x, y) = \frac{x'y}{\left[(x'x)(y'y)\right]^{1/2}}$$

EXPERIMENTAL RESULTS

Data Set

The proteins data used in this article were extracted from the data set used in the studies of Guo *et al*. [20]. The PPI data was collected from *S.cerevisiae* core subset of database of interacting proteins (DIP) [27], version DIP_20070219. After redundant proteins with fewer than 50 residues and proteins with sequence identity≥40% were removed, the date set contain 5594 PPIs. The 5594 non-interacting pairs were generated based on such an assumption that proteins occupying different subcellular localizations do not interact. All experiments are performed on this whole PPI data using 5-fold cross validation.

Evaluation Measure

To measure the performance of our method, we use 5-fold cross validation and the following criterion functions, where true positive (TP) is the number of true PPIs that are predicted correctly; true negative (TN) is the number of true non-interacting pairs that are predicted correctly; false positive (FP) is the number of true non-interacting pairs that are predicted to be PPIs; and false negative (FN) is the number of true PPIs that are predicted to be non-interacting pairs.

$$\textit{Accuracy}: \frac{\text{TN+TP}}{\text{TN+FP+FN+TP}}$$

$$\textit{Sensitivity}: \frac{\text{TP}}{\text{FN+TP}}$$

$$\textit{Precision}: \frac{\text{TN}}{\text{FP+TN}}$$

The Prediction Results

Selecting Optimal Parameters of KNNs

In our method, two parameters of KNNs, i.e. *k* and distance measuring function (*L1, L2* and *Cos*), are determined experimentally by using the first feature presentation of protein-protein pair (*Cod1*), which is defined above. Usually, *k* is selected as 1 or 3. Table **2** shows the results obtained with *Cod1* as the feature presentation of protein-protein pair and *L2* as the distance measuring function of KNNs. From the Table **2**, we can see that the predicting performance is slightly better when *k* is 3. Based on a fixed *k* and a kind of feature presentation of protein-protein pair, the question of which distance measuring function is more suitable in the context of PPIs prediction can be answered. Table **3** gives the results of KNNs prediction of interacting proteins with *Cod1* and *k*=3. It can be seen that the *L1* distance measuring function has an apparently advantage in the context of PPIs prediction. By the above experiments, two parameters of KNNs, *k* and distance measuring function (*L1, L2* and *Cos*), are selected as 3 and *L1*.

Comparing the Prediction Performances of Different Feature Presentations of Protein-Protein Pair

The predictive performances of four different feature presentations of protein-protein pair were compared and presented in Table **4**, using the optimized parameters of KNNs (*k*=3 and *L1*). It is found that *Cod4* has the best predictive performance in the context of PPIs prediction.

Base on the final optimized parameters (*k*=3, *L1* and *Cod4*), our method can achieved 86.15% prediction accuracy, 81.03% sensitivity and 90.24% precision when performed on the PPI data of yeast.

The final prediction model was built with the optimal parameters (*k*=3, *L1* and *Cod4*) and the whole *S.cerevisiae* PPI data (5594 PPIs).

Performance on the Independent Data Set

The real strength of any prediction method can be estimated only by evaluating its performance on an independent dataset. In this article, the PPI data collected from *E. coli* core subset of DIP [27], version DIP_20090126, was used to evaluate the practical prediction ability of the final prediction model. The *E. coli* data set containing 993 interaction pairs is the second largest in the species specific subsets of DIP, behind only the *S.cerevisiae.* After the protein pairs that contained a protein with <50 amino acids were removed, the remaining 986 protein pairs were performed by the final prediction model and the result shows that the prediction model can correctly predict 720 PPIs with 73.02% accuracy.

Table 2. The Comparative Results of the Prediction Performance Based on Different Values of *k*, Using *Cod1* as the Feature Presentation of Protein-Protein Pair and *L2* as the Distance Measuring Function

	***k*=1**	***k*=3**
Sensitivity(%)	68.48±1.20	68.76±1.49
Precision(%)	67.09±1.70	67.54±1.56
Accuracy(%)	67.43±0.88	67.84±1.12

Table 3. The Comparative Results of the Prediction Performance Based on Different Distance Measuring Functions, Using *Cod1* as the Feature Presentation of Protein-Protein Pair and Selecting *k* as 3

	L1	***L2***	***Cos***
Sensitivity(%)	75.81±1.20	68.76±1.49	66.82±1.74
Precision(%)	74.75±1.23	67.54±1.56	70.48±1.21
Accuracy(%)	75.08±1.13	67.84±1.12	69.40±1.21

Table 4. The Comparative Results of the Prediction Performance Based on Four Different Feature Presentations of Protein-Protein Pair, Using the Optimized Parameters of KNNs (*k*=3 and *L1*)

	Cod1	***Cod2***	***Cod3***	***Cod4***
Sensitivity(%)	75.81±1.20	76.77±0.69	78.14±0.90	81.03±1.74
Precision(%)	74.75±1.23	82.17±1.35	81.86±0.99	90.24±1.34
Accuracy(%)	75.08±1.13	80.04±1.06	80.41±0.47	86.15±1.17

The independent dataset from *E. coli* was also used to assess the performance of the combination between the proposed representation of local protein sequence descriptors and SVM, a widely used machine learning technique in the context of PPIs prediction. In this study, the software LIBSVM [28] was employed and the SVM models [29] were created with a set of optimal inbuilt parameters and kernels using *Cod4* as the feature presentation of protein-protein pair and the whole *S.cerevisiae* PPI data (5594 PPIs). In this case an accuracy of 71.24% was obtained. These results clearly indicate the stable performance of the proposed representation of local protein sequence descriptors when combining with different machine learning techniques.

CONCLUSIONS AND FUTURE WORK

With the large amount of protein sequences information provided by genome-sequencing projects, there is a growing demand for developing advanced computational methods for predicting potential PPIs by using sequence information only. Though some sequence-based methods are useful, one flaw of them is that the interactions information cannot be drawn from both continuous and discontinuous amino acids segments at the same time. To overcome this problem, in this paper we propose a sequence-based method based on a novel representation of local protein sequence descriptors. Local descriptors account for the interactions between residues in both continuous and discontinuous regions of a protein sequence, so this method enables us to draw more PPI information from the sequence. When performed on the PPI data of *S. cerevisiae*, the method achieved 86.15% prediction accuracy with 81.03% sensitivity at the precision of 90.24%. Meanwhile, the final prediction model was tested using the independent data set of the *E. coli* PPIs with a good performance. The performance of the combination between this novel representation of local protein sequence descriptors and SVM was also evaluated using the independent dataset from *E. coli* and the results indicate the stable performance of the proposed representation of local protein sequence descriptors when combining with different machine learning techniques. Given the complex nature of PPIs, the performance of our method is promising and it can be a helpful supplementary for PPIs prediction.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation of China, Nos. 60905023, 30900321 & 60975005, and the National Basic Research Program of China (973 Program), No. 2007CB311002, the grant of the Guide Project of Innovative Base of Chinese Academy of Sciences, No.KSCX1-YW-R-30, the Knowledge Innovation Program of the Chinese Academy of Sciences, No. 0823A16121 and the China Postdoctoral Science Foundation, No. 2009-0450707.

REFERENCES

[1] 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.

[2] Williams, N. E. Immunoprecipitation procedures. *Meth. Cell Biol.*, **2000**, *62*, 449-453.

[3] Zhu, H.; Bilgin, M.; Bangham, R.; Hall, D., Casamayor, A.; Bertone, P.; Lan, N.; Jansen, R.; Bidlingmaier, S. Global analysis of protein activities using proteome chips. *Science*, **2001**, *293*, 2101-2105.

[4] Uetz, P.; Giot, L.; Cagney, G.; Mansfield, T.A.; Judson, R.S.; Knight, J.R.; Lockshon, D.; Narayan, V.; Srinivasan, M.; Pochart, P.; Qureshi-Emili, A.; Li, Y.; Godwin, B.; Conover, D.; Kalbfleisch, T.; Vijayadamodar, G.; Yang, M.; Johnston, M.; Fields, S.; Rothberg, J.M. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, **2000**, *403*, 623-627.

[5] 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.

[6] Pazos, F.; Valencia, A. Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Eng.*, **2001**, *14*, 609-14.

[7] Wang, B.; Chen, P.; Huang, D.S.; Li, J.J.; Lok, T.M.; Lyu, M.R. Predicting protein interaction sites from residue spatial sequence profile and evolution rate. *FEBS Lett.*, **2006**, *580*(2), 380-384.

[8] Pazos, F.; Helmer-Citterich, M.; Ausiello, G.; Valencia, A. Correlated mutations contain information about protein-protein interaction. *J. Mol. Biol.*, **1997**, *271*, 511-523.

[9] Wang, B.; Wong, H.S.; Huang, D.S. Inferring protein-protein interacting sites using residue conservation and evolutionary information. *Protein Pept. Lett.*, **2006**, *13*(10), 999-1005.

[10] Li, J.J.; Huang, D.S.; Wang, B.; Chen, P. Identifying protein-protein interfacial residues in heterocomplexes using residue conservation scores. *Int. J. Biol. Macromol.*, **2006**, *38*, 241-247.

[11] Marcotte, E.M.; Xenarios, I.; Eisenberg, D. Mining literature for protein-protein interactions. *Bioinformatics*, **2001**, *17*, 359-363.

[12] Marcotte, E.M.; Pellegrini, M.; Thompson, M.J.; Yeates, T.O.; Eisenberg, D. A combined algorithm for genome-wide prediction of protein function. *Nature*, **1999,** *402*, 83-86.

[13] Chen, X.W.; Jeong, J.C. Sequence-based prediction of protein interaction sites with an integrative method. *Bioinformatics*, **2009**, *25*(5), 585-591.

[14] Bock, J.R.; Gough, D.A. Predicting protein--protein interactions from primary structure. *Bioinformatics*, **2001**, *17*, 455-460.

[15] Bock, J.R.; Gough, D.A. Whole-proteome interaction mining. *Bioinformatics*, **2003**, *19*, 125-134.

[16] Najafabadi, H.S.; Salavati, R. Sequence-based prediction of protein-protein interactions by means of codon usage. *Genome Biol.*, **2008**, *9*(5), R87.

[17] Liu, L.; Cai, Y.; Lu, W.; Feng, K.; Peng, C.; Niu, B. Prediction of protein-protein interactions based on PseAA composition and hybrid feature selection. *Biochem. Biophys. Res. Commun.*, **2009**, *380*, 318-322.

[18] 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*(1), 137-145.

[19] Shen, J.; Zhang, J.; Luo, X.; Zhu, W.; Yu, K.; Chen, K.; Li, Y.; Jiang, H. Predicting protein-protein interactions based only on sequences information. *Proc. Natl. Acad. Sci. USA*, **2007**, *104*(11), 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*(9), 3025-3030.

[21] Davies, M.N.; Secker, A.; Freitas, A.A.; Clark, E.; Timmis, J.; Flower, D.R. Optimizing amino acid groupings for GPCR classification. *Bioinformatics*, **2008**, *24*(18), 1980-1986.

[22] Tong, J.C.; Tammi, M.T. Prediction of protein allergenicity using local description of amino acid sequence. *Front. Biosci.*, **2008,** *13*, 6072-6078.

[23] Huang, D.S.; Zhao, X.M.; Huang, G.B.; Cheung, Y.M. Classifying protein sequences using hydropathy blocks. *Pattern Recogn.*, **2006**, *39*(12), 2293-2300.

[24] Zhao, X.M.; Cheung, Y.M.; Huang, D.S. A novel approach to extracting features from motif content and protein composition for protein sequence classification. *Neural Network.*, 2005, *18*(8), 1019-1028.

[25] Cui, J.; Han, L.Y.; Li, H.; Ung, C.Y.; Tang, Z.Q.; Zheng, C.J.; Cao, Z.W.; Chen, Y.Z. Computer prediction of allergen proteins from sequence-derived protein structural and physicochemical properties. *Mol. Immunol.*, **2007**, *44*, 514-520.

[26] Cover, T.M.; Hart, P.E. Nearest neighbor pattern classification. *IEEE Trans. Inform. Theor.*, **1967**, *13*, 21-27.

[27] Xenarios, I.; Salwinski, L.; Duan, X.J.; Higney, P.; Kim, S.M.; Eisenberg, D. DIP, the database of nteracting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.*, **2002**, *30*, 303-305.

[28] Chang, C.C.; Lin, C.J. LIBSVM: a Library for Support Vector Machines. Software available at: http://www. csie. ntu. edu. tw/cjlin/libsvm, **2001**.

[29] Xia, J.F.; Zhao, X.M.; Jiang, S.; Huang, D.S. APIS: accurate prediction of hot spots in protein interfaces by combining protrusion index with solvent accessibility. *BMC Bioinform.*, **2010**, *11*, 174.

Received: March 29, 2009 **Revised: May 15, 2009** **Accepted: June 02, 2009**