Biochemical and Biophysical Research Communications 417 (2012) 73–77

Contents lists available at SciVerse ScienceDirect

Biochemical and Biophysical Research Communications

journal homepage: www.elsevier.com/locate/ybbrc

# Prediction of pattern recognition receptor family using pseudo-amino acid composition

Qing-Bin Gao [a,b], Hongyu Zhao [b], Xiaofei Ye [a], Jia He [a,*]

[a] Department of Health Statistics, Second Military Medical University, Shanghai 200433, China
[b] School of Epidemiology and Public Health, Yale University, New Haven, CT 06520, USA

ARTICLE INFO

*Article history:*
Received 2 November 2011
Available online 19 November 2011

*Keywords:*
Pattern recognition receptor
Function prediction
Pseudo amino acid composition
Computational biology
Support vector machines

ABSTRACT

Pattern recognition receptors (PRRs) play a key role in the innate immune response by recognizing pathogen associated molecular patterns derived from a diverse collection of microbial pathogens. PRRs form a superfamily of proteins related to host health and disease. Thus, prediction of PRR family might supply biologically significant information for functional annotation of PRRs and development of novel drugs. In this paper, a computational method is proposed for predicting the families of PRRs. The prediction was performed on the basis of amino acid composition and pseudo-amino acid composition (PseAAC) from primary sequences of proteins using support vector machines. A non-redundant dataset consisted of 332 PRRs in seven families was constructed to do training and testing. It was demonstrated that different families of PRRs were quite closely correlated with amino acid composition as well as PseAAC. In the jackknife test, overall accuracies of amino acid composition-based and PseAAC-based classifiers reached 96.1% and 97.9%, respectively. The results indicate that families of PRRs are predictable with high accuracy. It is anticipated that this computational method might be a powerful tool for the automated assignment of families of PRRs.



## 1. Introduction

Germline-encoded pattern recognition receptors (PRRs) are a class of innate immune response-expressed proteins that are responsible for sensing the presence of microorganisms. They do this by recognizing structures conserved among microbial species, which are called pathogen-associated molecular patterns (PAMPs) [1]. Binding of microbial PAMPs to their PRRs promotes the synthesis and secretion of intracellular regulatory molecules, such as cytokines, that are crucial to initiating innate immunity and adaptive immunity. The study of PRR structure and function is essential for the proper understanding of their normal and abnormal cellular mechanisms because of both their physiology and pathophysiology. Research in the field of PRRs has grown enormously in the past few years and has broad relevance to multiple diseases. Since many PRRs are potential targets for clinical intervention and therapy, it is of interest both to researchers interested in signaling and also to the pharmaceutical and biotech industry. As a rising branch, automated prediction of families of PRRs is becoming crucial because the biological function of a receptor is closely correlated with its category.

Since the identification of Toll-like receptors, our knowledge about PRRs has increased rapidly. One alternative method for identifying additional members of PRRs is using sequence similarities to known PRRs, which may be performed by sequence similarity searching tools, such as BLAST [2] and FASTA [3]. However, one major limitation of these searching tools is that they are not able to recognize the families of PRRs. At present, PRRs have been classified into seven families according to the PRRDB database [4]. However, prediction of these families by using phylogeny or BLAST-based tools is difficult due to a scarcity of data for some families. Therefore, designing a reliable computational method to classify PRRs to their corresponding families automatically is strongly desired. With the accumulation of biological data generated by many large-scale genome sequencing projects, it is possible now to develop computational methods to predict the families of PRRs.

In this paper, an effective computational method was proposed for the first time for classifying the families of PRRs. The classification was performed on the basis of amino acid composition and various PseAAC of protein sequences. Amino acid composition is a simple approach for producing patterns of fixed length from protein sequences of varying length. PseAAC is now a very popular approach used for improving the prediction quality of diverse protein attributes by incorporating sequence order and structure information of proteins [5,6], such as subcellular localization [7–13], protein structural class [14–18], G-protein coupled receptors

* Corresponding author.
*E-mail address:* hejia63@yahoo.com (J. He).

0006-291X/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2011.11.057

[19–22], nuclear receptors [23], enzyme family class [24–27], protease type [28,29], protein folding rate [30], outer membrane protein [31,32], etc. [33–35]. We also examined the influence of rank of correlation factor and weighting factor (two parameters for computing PseAAC) on the prediction performance of the proposed method. In this study, support vector machines (SVMs) were used to construct classifiers. Different types of PseAAC were considered and the one with the best prediction quality was finally accepted to describe protein sequences.

## 2. Materials and methods

### 2.1. Dataset

The dataset of PRRs was extracted from database PRRDB available at http://www.imtech.res.in/raghava/prrdb/ [4]. PRRDB is a comprehensive database that collects information on PRRs and their ligands reported in literature. The current version of PRRDB contains seven known families of PRRs. By browsing different families of PRRs in PRRDB, users can obtain a unique Swiss-Prot ID assigned to each receptor. We used these Swiss-Prot IDs to retrieve protein sequences from UniProt Knowledgebase (UniProtKB) by submitting them as UniProt identifiers. As some sequences relevant to the Swiss-Prot IDs have been deleted from UniProtKB, the original dataset constructed in our study comprises 473 PRRs in seven families. To reduce the homology bias, a redundancy reduction procedure was performed on the original dataset. Sequences with a high degree of similarity to the other sequences in the dataset were removed by program CD-HIT [36,37]. To remove the homologous sequences from a dataset, a cutoff threshold of 25% was imposed in Refs. [38–40] to exclude those proteins from the datasets that have $\geqslant$25% sequence identity to any other proteins in the dataset. However, in this study we did not use such a stringent criterion because the currently available data do not allow us to do so. Otherwise, the number of proteins for some subsets would be too few to have statistical significance. We grouped all protein sequences by CD-HIT with the threshold of 0.9 to ensure that no sequence had $\geqslant$90% sequence similarity to any sequences in the dataset. After such a screening procedure, the resulting dataset contains 332 proteins belonging to seven families. The number of proteins in each family is shown in Table 1. This dataset is available on request from the authors.

### 2.2. Support vector machines

Support vector machine (SVM) is a popular machine learning algorithm based on structural risk minimization for pattern classification [41]. It has been widely used in the community of biological sequence analysis. The software adopted to implement SVMs is LibSVM [42]. The SVMs with RBF kernel were used to construct the classifiers. The SVM classifiers were trained with one-versus-rest method to handle the multiclass problem.

**Table 1**
The number of proteins in each family of the dataset.

| Family | No. of protein sequences |
|---|---|
| Toll-like | 106 |
| Scavenger | 80 |
| Nucleotide binding site-leucine repeats rich | 46 |
| Mannose receptors | 16 |
| C-type lectin like domain | 33 |
| Dendritic cell-specific ICAM-3-grabbing nonintegrin | 12 |
| PeptidoGlycan recognition protein | 39 |
| Total | 332 |

### 2.3. Sequence representation

To develop a classification model of SVMs, each protein sequence in the training dataset should be quantified by a feature vector, which is usually constituted by some protein features. In addition to the conventional amino acid composition, which has been widely used for encoding protein sequences, in the current study we attempt to use pseudo amino acid composition (PseAAC) of proteins to accomplish the prediction of PRR family. The web server PseAAC supplies us with a convenient tool for generating various types of PseAAC, which is freely available at http://www.chou.med.harvard.edu/bioinf/PseAA/ [43]. Three different parameters can be used to generate various kinds of PseAAC. They are quantitative characters of amino acids, rank of correlation and weighting factor. Now, six physicochemical characters of amino acids are supported to calculate the correlations between amino acids at different positions along protein sequence. They are hydrophobicity, hydrophilicity, side chain mass, pK of the $\alpha$-COOH group, pK of the $\alpha$-$NH_3^+$ group, and pI at 25 °C. Thus, 63 different parallel correlation types (type I) of PseAAC and 63 different series correlation types (type II) of PseAAC as well as the dipeptide PseAAC can be generated by PseAAC. The dimension for the output of type I PseAAC is $(20+\lambda)$ [6], dimension for the output of type II PseAAC is $(20+\xi\times\lambda)$ [44], and dimension for the dipeptide PseAAC is 420. Here, $\lambda$ is a non-negative integer smaller than the length of the input sequence representing the rank of correlation of amino acids along a protein sequence, and $\xi$ is the number of amino acid characters selected by the user. In our study, character of pI at 25 °C was not included for generating PseAAC, for it is a character at the specific temperature and led to a degeneration of prediction quality in our preliminary research, thus in this paper $\xi = 5$. This feature vector is expected to be able to encapsulate some sequence-order and structural information of proteins. Particularly, when $\lambda = 0$, PseAAC degenerates to the conventional amino acid composition. The weighting factor is designed for users to put weight on the additional PseAAC with respect to the conventional amino acid components. The prediction quality based on different parameters for PseAAC was discussed in this study. The values of each element of feature vector were normalized between 0 and 1 using the standard conversion formula before it was inputted into the prediction engine of SVMs.

### 2.4. Performance evaluation

In statistical prediction, independent dataset test, subsampling ($n$-fold cross-validation) test and jackknife test are often used to examine a predictor for its effectiveness in practical application. However, as elucidated by [39] and demonstrated in [45], among the three methods, jackknife test is deemed the most objective that can yield a unique result for a given dataset [46], and hence has been increasingly used by investigators to examine the accuracy of various predictors [47–51]. Accordingly, the jackknife test was adopted in this study to evaluate the quality of the proposed method. The performance metrics used for the evaluation of the classifiers are overall accuracy, accuracy and Matthews correlation coefficient (MCC). They are defined by [52]:

$$\text{Overall accuracy} = \frac{\sum_{i=1}^{k} p(i)}{N}, \tag{1}$$

$$\text{accuracy} = \frac{p(i)}{\text{obs}(i)}, \tag{2}$$

$$\text{MCC}(i) = \frac{p(i)n(i) - u(i)o(i)}{\sqrt{(p(i)+u(i))(p(i)+o(i))(n(i)+u(i))(n(i)+o(i))}} \tag{3}$$

where $N$ is the total number of sequences in the dataset, $k$ is the type number, obs($s$) is the number of sequences observed in family $i$, $p(i)$ is the number of correctly predicted sequences of family $i$, $n(i)$, is the number of correctly predicted sequences not in family $i$, $u(i)$ is the number of under-predicted sequences and $o(i)$ is the number of over-predicted sequences.

## 3. Results

To find an appropriate representation of protein sequences that could classify the families of PRRs with high accuracy, we investigated the prediction performance of type I and type II PseAAC, respectively. Type I and type II PseAAC were generated by web server PseAAC with different parameters. The PseAAC resulting in the highest prediction accuracy was accepted to represent protein sequences. The parameter $C$ for RBF kernel SVMs was set to 4 because $C = 4$ led to a higher accuracy in this study. The optimal values of parameter $\gamma$ for RBF kernel SVMs, which could lead to a better performance, were also provided in the resulting tables.

### 3.1. Results of type I PseAAC

In order to consider the effect of rank of correlation factor $\lambda$ on the prediction quality, we initially set weighting factor $w$ to 0.05. Different values of $\lambda$ (see Table 2) were then imposed to generate dimensionally different PseAAC for describing protein sequences. Table 2 shows the results of type I PseAAC with different $\lambda$ based on RBF kernel SVM classifiers. Since the classification accuracy of SVM classifier is dependent on parameter $\gamma$, Table 2 provided the highest accuracy with its corresponding $\gamma$. The optimal value of $\lambda$ is determined when it results in the highest overall accuracy. From Table 2 we can see that the overall accuracy for all $\lambda$ is $\geqslant$95%, indicating that the current classifier can reach a high accuracy. The results also demonstrate that the prediction performance is dependent on the value of $\lambda$. When $\lambda = 0$, PseAAC equals to conventional amino acid composition and its prediction accuracy reached 96.1%. For the current study, the optimal value of $\lambda$ is 40. Thus the dimension of the PseAAC considered here is 60 (20 + 40). The highest overall accuracy of 97.3% was observed in the jackknife test. The optimal value of $\lambda$ may vary either with the number and combination of the physicochemical characters selected by the user.

As a beneficial complement, we also examine the effect of weighting factor $w$ on the prediction quality, which adjusts the latter $\lambda$ components to be in similar scales with the first 20 amino acid composition components. During this procedure, $\lambda$ is fixed at 40, while $w$ is changeable. The prediction results with the jackknife test were shown in Table 3. From Table 3 we noticed that the $w$ had a slight influence on the overall accuracy. In the current condition, the best result was observed at $w = 0.05$ with the overall accuracy of 97.3%. This indicated that a smaller $w$ could lead to higher prediction accuracy.

**Table 2**
The overall accuracy of type I PseAAC with different $\lambda$ ($C = 4$).

| $\lambda$ | Dimension | $\gamma$ for RBF | Overall accuracy (%) |
|---|---|---|---|
| 0 | 20 | 3 | 96.1 |
| 5 | 25 | 3 | 95.8 |
| 10 | 30 | 2 | 95.8 |
| 15 | 35 | 2 | 95.5 |
| 20 | 40 | 1.0 | 96. 7 |
| 25 | 45 | 1.0 | 96.4 |
| 30 | 50 | 0.8 | 96.7 |
| 35 | 55 | 0.8 | 97.0 |
| **40** | **60** | **0.5** | **97.3** |
| 45 | 65 | 0.5 | 97.3 |
| 50 | 70 | 1.2 | 96.7 |
| 55 | 75 | 1.2 | 96.7 |
| 60 | 80 | 1.0 | 96.7 |
| 65 | 85 | 1.0 | 97.0 |
| 70 | 90 | 1.0 | 96.7 |
| 75 | 95 | 0.9 | 96.7 |
| 80 | 100 | 0.9 | 97.0 |

Highest accuracy is shown in bold.

**Table 3**
The overall accuracy of type I PseAAC with different $w$ ($\lambda = 40$).

| $w$ | $\gamma$ for RBF | Overall accuracy (%) |
|---|---|---|
| **0.05** | **0.5** | **97.3** |
| 0.1 | 0.5 | 97.0 |
| 0.15 | 0.5 | 97.0 |
| 0.2 | 0.8 | 97.0 |
| 0.25 | 0.8 | 97.0 |
| 0.3 | 0.8 | 97.0 |

Highest accuracy is shown in bold.

**Table 4**
The overall accuracy of type II PseAAC with different $\lambda$ ($C = 4$).

| $\lambda$ | Dimension | $\gamma$ for RBF | Overall accuracy (%) |
|---|---|---|---|
| 5 | 45 | 0.9 | 97.6 |
| 10 | 70 | 0.6 | 96.7 |
| 15 | 95 | 0.3 | 97.3 |
| 20 | 120 | 0.2 | 97.3 |
| 25 | 145 | 0.25 | 97.6 |
| 30 | 170 | 0.25 | 97.6 |
| **35** | **195** | **0.2** | **97.9** |
| 40 | 220 | 0.2 | 97.9 |
| 45 | 245 | 0.15 | 97.3 |
| 50 | 270 | 0.25 | 96.7 |

Highest accuracy is shown in bold.

### 3.2. Results of type II PseAAC

Similarly, we first set weighting factor $w$ to 0.05 to investigate the relationship between the prediction quality and the rank of correlation factor $\lambda$. Table 4 shows the results of present method based on type II PseAAC with different $\lambda$. We noticed that a high accuracy of 97.9% was reached in the jackknife test when $\lambda = 35$, and the dimension of the PseAAC considered is 195 ($20 + 5 \times 35$). The result indicates that prediction accuracy of type II PseAAC is a little bit higher than that of type I PseAAC. Therefore, type II PseAAC is a preferred choice of representation for distinguishing the families of PRRs. As to the effect of weighting factor $w$ on the prediction quality of type II PseAAC, the conclusion is similar to that of type I PseAAC, as shown in Table 5. In this study, the best result was also observed at $w = 0.05$ with the overall accuracy of 97.9% by the jackknife test.

From Table 4 we also noticed that although the overall accuracy of $\lambda = 35$ was higher than that of $\lambda = 5$, it has more features (195 versus 45). Generally, fitting the data with a minimum number of features may increase the robustness of the results and decrease

**Table 5**
The overall accuracy of type II PseAAC with different $w$ ($\lambda = 35$).

| $w$ | $\gamma$ for RBF | Overall accuracy (%) |
|---|---|---|
| **0.05** | **0.2** | **97.9** |
| 0.1 | 0.2 | 97.6 |
| 0.15 | 0.2 | 97.6 |
| 0.2 | 0.2 | 97.6 |
| 0.25 | 0.3 | 97.0 |
| 0.3 | 0.7 | 96.7 |

Highest accuracy is shown in bold.

**Table 6**
The overall accuracy of type II PseAAC with different $w$ ($\lambda = 5$).

| $w$ | $\gamma$ for RBF | Overall accuracy (%) |
|---|---|---|
| **0.05** | **0.9** | **97.6** |
| 0.1 | 0.9 | 97.6 |
| 0.15 | 0.9 | 97.6 |
| 0.2 | 0.9 | 97.3 |
| 0.25 | 0.9 | 97.0 |
| 0.3 | 0.9 | 97.0 |

Highest accuracy is shown in bold.

**Table 7**
Comparison of type I PseAAC with type II PseAAC.

| Family | Type I ($\lambda = 40$, 60 D) | | Type II ($\lambda = 5$, 45 D) | |
|---|---|---|---|---|
| | Acc (%) | MCC | Acc (%) | MCC |
| Toll-like | 97.2 | 0.98 | 97.2 | 0.98 |
| Scavenger | 98.8 | 0.95 | 100 | 0.97 |
| NBSL repeats rich | 100 | 0.99 | 100 | 0.99 |
| Mannose receptors | 100 | 1.00 | 100 | 1.00 |
| C-type lectin like | 93.9 | 0.93 | 93.9 | 0.95 |
| Dendritic cell-specific | 91.7 | 0.96 | 91.7 | 0.96 |
| PeptidoGlycan | 94.9 | 0.96 | 94.9 | 0.94 |
| Overall | 97.3 | – | 97.6 | – |

Acc: accuracy.

the curse of dimensionality as well as the risk of overfitting. Therefore, in the present study we preferred the 45 features ($\lambda = 5$) giving the accuracy of 97.6%. This is almost as high as the accuracy of 97.9% obtained by using the 195 features ($\lambda = 35$). Table 6 shows the influence of weighting factor $w$ on the prediction quality of type II PseAAC when $\lambda = 5$. The highest accuracy is also observed at $w = 0.05$, which is consensus to what we have obtained in the above section. Thus, in this work type II PseAAC generated with $\lambda = 5$ and $w = 0.05$ was recommended to predict the families of PRRs.

Furthermore, we also made a comparison between type I and type II PseAAC on the prediction quality in each family, as shown in Table 7. The results indicated that type II PseAAC achieved a similar or even higher accuracy in all families of PRRs with fewer features. However, when individual families were examined the prediction quality for some small groups became worse, such as C-type lectin like and Dendritic cell-specific (Table 7). In the future, further improvements probably can be obtained by preparing large dataset with higher quality. It might be possible to increase the number of data entries from updated database, especially for those small groups with low prediction accuracy. This is because the performance of any knowledge based method is dependent on the quality and quantity of the data.

## 4. Discussion

In this paper, a computational method was proposed for the first time for predicting the families of PRRs using protein primary sequences. Five physicochemical characters of amino acids were used to generate the sequence features of PseAAC. Different values of rank of correlation factor and weighting factor for generating PseAAC were tested to find the most appropriate representation of proteins. Finally, $\lambda = 40$ and $w = 0.05$ were selected to generate the feature vector of type I PseAAC, while $\lambda = 5$ and $w = 0.05$ were selected to generate the feature vector of types II PseAAC. The overall accuracies of 97.3% and 97.6% were observed in the jackknife test based on the produced type I and type II PseAAC, respectively. The study proves that there is a close correlation between the protein features (amino acid composition and PseAAC) and the families of PRRs. According to the results we know that type II PseAAC is a better representation of proteins for the prediction of PRRs. The development of this novel method can serve as an effective tool for predicting the families of PRRs and thus will speed up the pace of family identification of new PRRs.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 30901243), the Natural Science Foundation of Shanghai (No. 09ZR1438900) and the Chinese Key Project for Infectious Diseases (2008ZX10002-018).

## References

[1] O. Takeuchi, S. Akira, Pattern recognition receptors and inflammation, Cell 140 (2010) 805–820.
[2] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410.
[3] W.R. Pearson, D.J. Lipman, Improved tools for biological sequence comparison, Proc. Natl. Acad. Sci. U S A 85 (1988) 2444–2448.
[4] S. Lata, G.P. Raghava, PRRDB: a comprehensive database of pattern-recognition receptors and their ligands, BMC Genomics 9 (2008) 180.
[5] K.C. Chou, Some remarks on protein attribute prediction and pseudo amino acid composition, J. Theor. Biol. 273 (2011) 236–247.
[6] K.C. Chou, Prediction of protein cellular attributes using pseudo-amino acid composition, Proteins 43 (2001) 246–255.
[7] X. Yu, X. Zheng, T. Liu, Y. Dou, J. Wang, Predicting subcellular location of apoptosis proteins with pseudo amino acid composition: approach from amino acid substitution matrix and auto covariance transformation, Amino Acids (2011), doi:10.1007/s00726-011-0848-8.
[8] R. Shi, C. Xu, Prediction of rat protein subcellular localization with pseudo amino acid composition based on multiple sequential features, Protein Peptide Lett. 18 (2011) 625–633.
[9] T. Liu, X. Zheng, C. Wang, J. Wang, Prediction of subcellular location of apoptosis proteins using pseudo amino acid composition: an approach from auto covariance transformation, Protein Peptide Lett. 17 (2010) 1263–1269.
[10] K.K. Kandaswamy, G. Pugalenthi, S. Moller, E. Hartmann, K.U. Kalies, P.N. Suganthan, T. Martinetz, Prediction of apoptosis protein locations with genetic algorithms and support vector machines through a new mode of pseudo amino acid composition, Protein Peptide Lett. 17 (2010) 1473–1479.
[11] J.Y. Shi, S.W. Zhang, Q. Pan, G.P. Zhou, Using pseudo amino acid composition to predict protein subcellular location: approached with amino acid composition distribution, Amino Acids 35 (2008) 321–327.
[12] F.M. Li, Q.Z. Li, Using pseudo amino acid composition to predict protein subnuclear location with improved hybrid approach, Amino Acids 34 (2008) 119–125.
[13] K.C. Chou, Y.D. Cai, Prediction and classification of protein subcellular location-sequence-order effect and pseudo amino acid composition, J. Cell. Biochem. 90 (2003) 1250–1260.
[14] J. Wu, M.L. Li, L.Z. Yu, C. Wang, An ensemble classifier of support vector machines used to predict protein structural classes by fusing auto covariance and pseudo-amino acid composition, Protein J. 29 (2010) 62–67.
[15] S.S. Sahu, G. Panda, A novel feature representation method based on Chou's pseudo amino acid composition for protein structural class prediction, Comput. Biol. Chem. 34 (2010) 320–327.
[16] Z.C. Li, X.B. Zhou, Z. Dai, X.Y. Zou, Prediction of protein structural classes by Chou's pseudo amino acid composition: approached using continuous wavelet transform and principal component analysis, Amino Acids 37 (2009) 415–425.
[17] T.L. Zhang, Y.S. Ding, K.C. Chou, Prediction protein structural classes with pseudo-amino acid composition: approximate entropy and hydrophobicity pattern, J. Theor. Biol. 250 (2008) 186–193.
[18] X. Xiao, P. Wang, K.C. Chou, Predicting protein structural classes with pseudo amino acid composition: an approach using geometric moments of cellular automaton image, J. Theor. Biol. 254 (2008) 691–696.
[19] Z. Ur-Rehman, A. Khan, G-protein-coupled receptor prediction using pseudo-amino-acid composition and multiscale energy representation of different physiochemical properties, Anal. Biochem. 412 (2011) 173–182.
[20] Q. Gu, Y.S. Ding, T.L. Zhang, Prediction of G-protein-coupled receptor classes in low homology using Chou's pseudo amino acid composition with approximate entropy and hydrophobicity patterns, Protein Peptide Lett. 17 (2010) 559–567.
[21] J.D. Qiu, J.H. Huang, R.P. Liang, X.Q. Lu, Prediction of G-protein-coupled receptor classes based on the concept of Chou's pseudo amino acid composition: an approach from discrete wavelet transform, Anal. Biochem. 390 (2009) 68–73.
[22] R. Zia Ur, A. Khan, Prediction of GPCRs with pseudo amino acid composition: employing composite features and grey incidence degree based classification, Protein Peptide Lett. 18 (2011) 872–878.
[23] Q.B. Gao, Z.C. Jin, X.F. Ye, C. Wu, J. He, Prediction of nuclear receptors with optimal pseudo amino acid composition, Anal. Biochem. 387 (2009) 54–59.

[24] Y.C. Wang, X.B. Wang, Z.X. Yang, N.Y. Deng, Prediction of enzyme subfamily class via pseudo amino acid composition by incorporating the conjoint triad feature, Protein Peptide Lett. 17 (2010) 1441–1449.
[25] J.D. Qiu, J.H. Huang, S.P. Shi, R.P. Liang, Using the concept of Chou's pseudo amino acid composition to predict enzyme family classes: an approach with support vector machine based on discrete wavelet transform, Protein Peptide Lett. 17 (2010) 715–722.
[26] X.B. Zhou, C. Chen, Z.C. Li, X.Y. Zou, Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes, J. Theor. Biol. 248 (2007) 546–551.
[27] Y.D. Cai, K.C. Chou, Predicting enzyme subclass by functional domain composition and pseudo amino acid composition, J. Proteome Res. 4 (2005) 967–971.
[28] L. Hu, L. Zheng, Z. Wang, B. Li, L. Liu, Using pseudo amino acid composition to predict protease families by incorporating a series of protein biological features, Protein Peptide Lett. 18 (2011) 552–558.
[29] G.P. Zhou, Y.D. Cai, Predicting protease types by hybridizing gene ontology and pseudo amino acid composition, Proteins 63 (2006) 681–684.
[30] J. Guo, N. Rao, G. Liu, Y. Yang, G. Wang, Predicting protein folding rates using the concept of Chou's pseudo amino acid composition, J. Comput. Chem. 32 (2011) 1612–1617.
[31] Q.B. Gao, X.F. Ye, Z.C. Jin, J. He, Improving discrimination of outer membrane proteins by fusing different forms of pseudo amino acid composition, Anal. Biochem. 398 (2010) 52–59.
[32] H. Lin, The modified Mahalanobis discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition, J. Theor. Biol. 252 (2008) 350–356.
[33] J.D. Qiu, S.B. Suo, X.Y. Sun, S.P. Shi, R.P. Liang, OligoPred: a web-server for predicting homo-oligomeric proteins by incorporating discrete wavelet transform into Chou's pseudo amino acid composition, J. Mol. Graph. Model. 30 (2011) 129–134.
[34] H. Mohabatkar, M. Mohammad Beigi, A. Esmaeili, Prediction of GABAA receptor proteins using the concept of Chou's pseudo-amino acid composition and support vector machine, J. Theor. Biol. 281 (2011) 18–23.
[35] D. Wang, L. Yang, Z. Fu, J. Xia, Prediction of thermophilic protein with pseudo amino acid composition: an approach from combined feature selection and reduction, Protein Peptide Lett. 18 (2011) 684–689.
[36] W. Li, L. Jaroszewski, A. Godzik, Clustering of highly homologous sequences to reduce the size of large protein databases, Bioinformatics 17 (2001) 282–283.
[37] W. Li, A. Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, Bioinformatics 22 (2006) 1658–1659.
[38] H.B. Shen, J. Yang, K.C. Chou, Euk-PLoc: an ensemble classifier for large-scale eukaryotic protein subcellular location prediction, Amino Acids 33 (2007) 57–67.
[39] K.C. Chou, H.B. Shen, Cell-PLoc: a package of web servers for predicting subcellular localization of proteins in various organisms, Nat. Protoc. 3 (2008) 153–162.
[40] H.B. Shen, K.C. Chou, Identification of proteases and their types, Anal. Biochem. 385 (2009) 153–160.
[41] V. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.
[42] C.C. Chang, C.J. Lin, LIBSVM: a library for support vector machines, Software is available from http://www.csie.ntu.edu.tw/~cjlin/libsvm.
[43] H.B. Shen, K.C. Chou, PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition, Anal. Biochem. 373 (2008) 386–388.
[44] K.C. Chou, Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes, Bioinformatics 21 (2005) 10–19.
[45] K.C. Chou, H.B. Shen, Recent progress in protein subcellular location prediction, Anal. Biochem. 370 (2007) 1–16.
[46] K.C. Chou, C.T. Zhang, Prediction of protein structural classes, Crit. Rev. Biochem. Mol. Biol. 30 (1995) 275–349.
[47] C. Chen, X. Zhou, Y. Tian, X. Zou, P. Cai, Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network, Anal. Biochem. 357 (2006) 116–121.
[48] P. Du, Y. Li, Prediction of protein submitochondria locations by hybridizing pseudo-amino acid composition with various physicochemical features of segmented sequence, BMC Bioinf. 7 (2006) 518.
[49] Q.B. Gao, Z.Z. Wang, Classification of G-protein coupled receptors at four levels, Protein Eng. Des. Sel. 19 (2006) 511–516.
[50] H. Lin, Q.Z. Li, Predicting conotoxin superfamily and family by using pseudo amino acid composition and modified Mahalanobis discriminant, Biochem. Biophys. Res. Commun. 354 (2007) 548–551.
[51] Y. Fang, Y. Guo, Y. Feng, M. Li, Predicting DNA-binding proteins: approached from Chou's pseudo amino acid composition and other specific sequence features, Amino Acids 34 (2008) 103–109.
[52] S. Hua, Z. Sun, Support vector machine approach for protein subcellular localization prediction, Bioinformatics 17 (2001) 721–728.