Amino Acids (2007) 33: 423–428
DOI 10.1007/s00726-006-0485-9
Printed in The Netherlands

Amino Acids

Prediction of linear B-cell epitopes using amino acid pair antigenicity scale*

J. Chen[1], **H. Liu**[1], **J. Yang**[1], and **K.-C. Chou**[1,2]

[1] Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong University, Shanghai, China
[2] Gordon Life Science Institute, San Diego, CA, U.S.A.

Received October 15, 2006
Accepted November 28, 2006
Published online January 26, 2007;

Summary. Identification of antigenic sites on proteins is of vital importance for developing synthetic peptide vaccines, immunodiagnostic tests and antibody production. Currently, most of the prediction algorithms rely on amino acid propensity scales using a sliding window approach. These methods are oversimplified and yield poor predicted results in practice. In this paper, a novel scale, called the amino acid pair (AAP) antigenicity scale, is proposed that is based on the finding that B-cell epitopes favor particular AAPs. It is demonstrated that, using SVM (support vector machine) classifier, the AAP antigenicity scale approach has much better performance than the existing scales based on the single amino acid propensity. The AAP antigenicity scale can reflect some special sequence-coupled feature in the B-cell epitopes, which is the essence why the new approach is superior to the existing ones. It is anticipated that with the continuous increase of the known epitope data, the power of the AAP antigenicity scale approach will be further enhanced.

Keywords: B-cell epitope – AAP antigenicity scale – SVM classifier

Abbreviations: AAP, amino acid pair; SVM, support vector machine; ROC, receiver operating characteristics

1. Introduction

An epitope is the part of a macromolecule that is recognized by the immune system, specifically by antibodies, cytotoxic T cells, or B cells. Although epitopes are usually thought to be derived from nonself proteins, sequences derived from the host that can be recognized are also classified as epitopes.

Located on the surface of the antigen, the B-cell epitopes are antigenic determinants recognized and bound by the B-cell receptor. B-cell epitopes are classified as either linear or discontinuous epitopes: the former comprises a single continuous stretch of amino acids within a protein sequence; while the latter consists of residues that are distantly separated in the sequence but are brought into physical proximity via protein folding. So far most experimental epitope detections have been focused on the linear B-cell epitopes. For simplicity, hereafter the B-cell epitopes will just refer to the linear ones.

Identification of epitopes on proteins is of vital importance for developing synthetic peptide vaccines, immunodiagnostic tests and antibody production. Most of the previous prediction methods in this regard were based on a single physicochemical property (or propensity scale) of the constituent amino acids, such as tight turns (Chou and Fasman, 1978), surface accessibility (Emini et al., 1985), flexibility (Karplus and Schulz, 1985), hydrophobicity (Parker et al., 1986), and antigenicity (Kolaskar and Tongaonkar, 1990). Currently, there are a set of linear B-cell prediction softwares available on the web, such as PREDITOP (Kolaskar and Tongaonkar, 1990), PEOPLE (Alix, 1999), and BEPITOPE (Odorico and Pellequer, 2003). All these algorithms have a common feature, i.e., calculating the average amino acid propensity value over a sliding window along a query protein sequence. The peak of the resulting profile is considered to correspond to the location of the B-cell epitope for the protein concerned. Unfortunately, to one's disappointment, a recent study by Blythe and Flower (2005) led to the conclusion that ''single-scale amino acid propensity profiles cannot be used to predict epitope location reliably'', as reflected by the fact that even the best amino acid propensity scales could only yield a success rate marginally better than that by randomly using ROC (receiver operating characteristics)

* Electronic supplementary material: Supplementary material is available in the online version of this article at http://dx.doi.org/10.1007/s00726-006-0485-9 and is accessible for authorised users.

plot (Delacour et al., 2005). As the data of linear B-cell epitopes are accumulating, it is possible to increase the prediction quality by using machine learning approaches (Sollner, 2006; Sollner and Mayer, 2006), and the prediction accuracy has been improved by so doing. Yet, it would not change the key issue, i.e., how to extract appropriate features from the primary sequence of B-cell epitopes no matter what kind of machine learning approach is used.

Stimulated by the facts that it can significantly enhance the success rates if the sequence coupling effects can be effectively taken into account as observed in the predictions of beta-turns (Chou, 1997b; Chou and Blinn, 1997), alpha-turns (Chou, 1997a), HIV-protease cleavage sites in proteins (Chou, 1993; Chou and Zhang, 1993; Zhang and Chou, 1993), specificity of GalNAc-transferase (Chou, 1995), signal peptides (Chou, 2001a, b; Liu et al., 2005), as well as the reports in three relevant reviews (Chou, 1996, 2000, 2002), we are to use the simplest amino acid coupling mode, i.e. the amino acid pairs (AAPs), to improve the prediction quality for the B-cell epitopes. The AAP approach has been successfully used to enhance the success rates for predicting the secondary structural contents in a protein (Chou, 1999; Liu and Chou, 1999). Actually, it is indicated through a statistical analysis on the epitope and non-epitope data sets that some particular AAPs are significantly higher than the others in the occurrence frequency. The present study was initiated in an attempt to use this kind of coupling effects to improve the prediction quality for B-cell epitopes.

2. Materials and methods

2.1 Data sets

The B-cell epitope data set was taken from Bcipep database (Saha et al., 2005), which is a collection of experimentally determined B-cell epitopes. It comprises 2479 continuous epitopes from over 1000 antigenic proteins. To train any learning machine, one needs to fix the length pattern. However, the B-cell epitopes are various in lengths. In order to create fixed length pattern, we adopted a ''truncation-extension treatment''. According to such an approach, if a B-cell epitope is longer than 20 amino acids, then only its central 20 residues are kept by equally truncating the surplus residues at both the N- and C-terminals; if a B-cell epitope is shorter than 20 amino acids, then the peptide segment is equally extended toward both the N- and C-terminals along the protein chain until it reach 20 residues long. If there were several identical amino acid segments after such a truncation-extension treatment, only one of them was kept. The epitope data set, also called positive data set $\mathbb{S}^+$, thus derived from the original Bcipep database (Saha et al., 2005) contains 872 unique epitopes with a same length of 20 residues. The non-epitope data set, also called negative dataset $\mathbb{S}^-$, was generated by randomly picking the 872 peptide segments of 20 amino acids from the Swiss-Prot database with the criterion that none of them is the same with any one of the segments in $\mathbb{S}^+$. Thus, we have a positive dataset $\mathbb{S}^+$ consisting of 872 B-cell epitopes, and a negative dataset $\mathbb{S}^-$ consisting of 872 non-B-cell epitopes (see Online Supporting Information A and B, respectively). They are the same in length but different in amino acid components and sequence order.

Table 1. Difference of AAP composition in Bcipep and Swiss-Prot database

AAP	Bcipep (%)	Swiss-Prot (%)	Ratio
WG	0.276	0.077	3.563
PG	0.956	0.369	2.591
WK	0.202	0.069	2.911
MP	0.037	0.111	0.331
FI	0.067	0.242	0.278

2.2 AAP Antigenicity scale

It is shown through a comparison of the Bcipep database with the Swiss-Prot database that the AAP composition for the epitopes is very different from that of the non-epitopes. AAPs are generated by decomposing the peptides of protiens continuously. For example, the peptide AEACCGCA can be decomposed into 7 AAPs: AE, EA, AC, CC, CG, GC, and CA. Listed in Table 1 are the occurrence frequencies of some AAPs, which indicate a great discrepancy between the epitopes and non-epitopes. If a query peptide segment contains the AAPs that epitopes prefer, its chance to be an epitope is higher, and vice versa. In order to capture this information, we introduce the ''AAP antigenicity scale'' for every AAP that can be interpreted as the likelihood of this AAP being associated with an epitope. There are $20 \times 20 = 400$ AAPs (Liu and Chou, 1999), so the AAP antigenicity scale also contains 400 entries. The AAP antigenicity scale is defined by

$$R_{\mathrm{AAP}} = \log\left(\frac{f^{+}_{\mathrm{AAP}}}{f^{-}_{\mathrm{AAP}}}\right) \tag{1}$$

where f^{+}_{AAP} and f^{-}_{AAP} are the occurrence frequencies of a given AAP in the epitopes and non-epitopes, respectively, and they can be derived from the Bcipep database and Swiss-Prot database, respectively. The AAP antigenicity scale (1) is normalized into $[-1, +1]$ through the following conversion:

$$R_{\mathrm{AAP}} = 2\left(\frac{R_{\mathrm{AAP}} - \min}{\max - \min}\right) - 1 \tag{2}$$

where max and min represent the maximum and minimum values among all the possible R_{AAP} values before the normalization. The normalization procedure of (2) will avoid the dominance of an individual feature in the classifier learning.

2.3 The SVM classifier and its parameter selection

In the previous prediction methods, the sliding windows along the primary sequence of a query protein was used and the propensity value of the central amino acid for the sequence highlighted by the window was adopted as a criterion in identifying whether the segment is a B-cell epitope. Obviously, such an approach is oversimplified. Also, it is quite arbitrary in choosing the threshold value. In this paper, the support vector machine (SVM) (Vapnik, 1998) was used as the prediction engine. In SVM each sample is represented by a vector. The SVM classifier basically learns how to classify a query sample $\boldsymbol{x}$ into two classes $\{-1, +1\}$ from a set of labeled training examples $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$. It derives a decision boundary between the positive and negative data sets with the maximum margin. Here, every component of the input vector $\boldsymbol{x}$ is a feature of the

peptide. The AAP antigenicity component of $\boldsymbol{x}$ is the average of the AAP antigenicity values of all the AAPs in the peptide.

When performing the classification with the SVM classifier, the choice of the penalty factor C and the type of kernel is very important. In this paper, if $\boldsymbol{x}$ consists of only one component, the dot kernel is used and the penalty parameter C used to control the trade-off between the errors of the SVM on training data and the margin maximization is selected from the range [10, 200] with the interval of 10; if $\boldsymbol{x}$ consists multiple components, i.e., a combination of different features of a peptide, the RBF kernel is used. The RBF is by far the most popular choice of kernel type used in SVM for its localization property. It is defined as (Scholkopf et al., 1997):

$$K(\boldsymbol{x},\boldsymbol{x}') = \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|^2}{2\sigma^2}\right) \tag{3}$$

where σ^2 is a control parameter reflecting the kernel width. To identify the optimal C and σ^2, a systematic grid search was conducted for C among $(2^{-2}, 2^{-1}, 2^0, 2^1 \ldots, 2^{11}, 2^{12})$ and for σ^2 among $(2^{-5}, \ldots, 2^{-1}, 2^0, 2^1, \ldots, 2^9)$. So there were $15\times15 = 225$ combinations. Finally, it was found that the optimal values for the two parameters are $C = 2^5 = 32$ and $\sigma^2 = 2$, and they were used for performing the prediction.

2.4 Performance measure

In this study, both the five-fold and jackknife cross-validations have been used. The five-fold cross-validation belongs to the sub-sampling test method. In the five-fold cross-validation, the data set is randomly divided into five subsets, each containing nearly equal number of peptides. The five subsets are then grouped into a training set and a testing set. The training set consists of four of these subsets and the testing set consists of the remaining one. This procedure is repeated five times and every subset is used once for testing. The final prediction results are the average of the five testing results.

In the jackknife test, each protein in the benchmark dataset was singled out in turn as a ''test protein'' and all the rule parameters were calculated from the remaining proteins. During the process of jackknife test, both the training and testing datasets were actually open, and a protein was in turn moving from one to the other. As is known, the single independent dataset test, sub-sampling test and jackknife test are the three methods often used for cross-validation in statistical prediction. Of these three, however, the jackknife test is deemed as the most rigorous and objective one, as illustrated by a comprehensive review (Chou and Zhang, 1995). Therefore, the jackknife test has been recently increasingly used in literatures (Cao et al., 2006; Chen et al., 2006; Chou and Shen, 2006a, b; Feng, 2001; Gao et al., 2005a, b; Guo et al., 2006; Sun and Huang, 2006; Wen et al., 2006; Xiao et al., 2005; Zhang et al., 2006; Zhou, 1998; Zhou and Doctor, 2003) for examining the power of various prediction methods.

In this study, the default threshold for SVM classifier is set to zero and all the results are given under this threshold. The following equations were usually used to measure the prediction quality:

$$\begin{cases} \Re_{\text{sen}} = \dfrac{\text{TP}}{\text{TP}+\text{FN}} \times 100\% \\ \Re_{\text{spe}} = \dfrac{\text{TN}}{\text{TN}+\text{FP}} \times 100\% \\ \Re_{\text{aac}} = \dfrac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{TN}+\text{FN}} \times 100\% \\ \Re_{\text{pos}} = \dfrac{\text{TP}}{\text{TP}+\text{FP}} \end{cases} \tag{4}$$

where $\Re_{\text{sen}}$ reflects the sensitivity, i.e., the rate of epitopes that are correctly predicted as epitopes; $\Re_{\text{spe}}$ reflects the specificity, i.e., is the rate of non-epitopes correctly predicted as non-epitopes; $\Re_{\text{acc}}$ reflects the accuracy, i.e., the rate of correctly predicted peptides; $\Re_{\text{pos}}$ reflects the positive prediction rate, i.e., the rate that a predicted epitope is in fact an epitope; TP represents the true positive; TN, the true negative; FP, the false positive; and FN, the false negative (Fig. 1). Also, the Matthew's correlation coefficient is computed according to the following formulation:

$$\text{MCC} = \frac{(\text{TP})(\text{TN}) - (\text{FP})(\text{FN})}{\sqrt{[\text{TP}+\text{FP}][\text{TP}+\text{FN}][\text{TN}+\text{FP}][\text{TN}+\text{FN}]}} \tag{5}$$

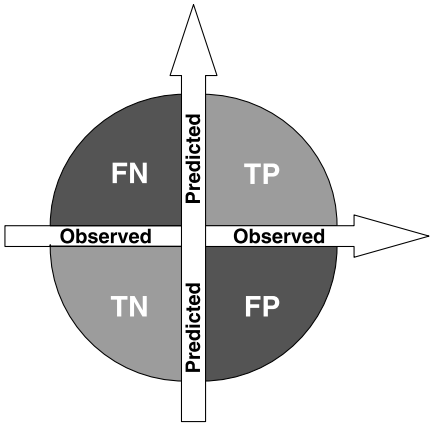

Fig. 1. An illustration to show: TP (true positive), the number of correct predictions for the positive dataset $\mathbb{S}^+$; TN (true negative), the number of correct predictions for the negative dataset $\mathbb{S}^-$; FP (false positive), the number of incorrect predictions for the positive dataset $\mathbb{S}^+$; and FN (false negative), the number of incorrect predictions for the negative dataset $\mathbb{S}^-$

Table 2. The performance of SVM using AAP antigenicity scale with different window sizes[a]

Window size	Sensitivity ($\Re_{\text{sen}}$)	Specificity ($\Re_{\text{spe}}$)	Accuracy ($\Re_{\text{acc}}$)	Positive prediction rate ($\Re_{\text{pos}}$)	Matthew correlation coefficient (MCC)
10	59.60	69.26	65.87	64.39	0.2902
12	58.81	71.05	66.79	64.93	0.3001
14	59.72	75.55	71.11	67.54	0.3571
16	60.93	74.48	70.80	67.62	0.3574
18	60.63	75.10	70.52	67.92	0.3616
20	60.87	75.36	71.09	68.07	0.3663

[a] These results were obtained by using SVM with the dot kernel and the penalty parameter C was optimized over the range of [10, 200] at the interval 10

Table 3. Comparison of performance of SVM using different scales and their combination by the 5-fold cross-validation[a]

Scale	Sensitivity ($\Re_{sen}$)	Specificity ($\Re_{spe}$)	Accuracy ($\Re_{acc}$)	Positive prediction rate ($\Re_{pos}$)	Matthew correlation coefficient (MCC)
Turns	53.32	65.48	60.44	59.31	0.1901
Accessibility	58.65	53.39	55.09	55.80	0.1211
Antigenicity	58.31	54.59	56.41	56.28	0.1295
Hydrophilicity	59.33	59.16	60.22	59.13	0.1852
Flexibility	57.66	57.60	57.59	57.47	0.1530
AAP antigenicity	60.87	75.36	71.09	68.07	0.3663
Combination of all	63.56	76.48	72.54	70.06	0.4040

[a] These results were obtained with the window size set at 20. The relevant parameters used for SVM in the combination method were $C = 32$ and $\sigma^2 = 2$

Table 4. Comparison of performance of SVM using different scales and their combination by the jackknife cross-validation[a]

Scale	Sensitivity ($\Re_{sen}$)	Specificity ($\Re_{spe}$)	Accuracy ($\Re_{acc}$)	Positive prediction rate ($\Re_{pos}$)	Matthew correlation coefficient (MCC)
Turns	51.61	65.71	60.08	58.66	0.1749
Accessibility	56.08	52.87	54.33	54.47	0.0895
Antigenicity	57.45	53.67	55.36	55.56	0.1113
Hydrophilicity	60.67	57.11	58.58	58.89	0.1779
Flexibility	54.82	58.03	56.64	56.42	0.1285
AAP antigenicity	58.94	74.31	69.65	66.63	0.3366
Combination of all	62.04	77.87	73.71	69.95	0.4042

[a] See footnote of Table 3 for further explanation

In addition to use of the threshold-dependent measure, we have also compared the sensitivity and specificity across all the possible thresholds to generate a ROC plot (Delacour et al., 2005). The ROC plot is a threshold-independent measure which can evaluate the overall performance of a prediction method. It was obtained by plotting all the sensitivity values (true positive rate) against their equivalent (1-specificity) values (false positive rate). The area under the ROC curve is deemed as an important measure of the overall prediction accuracy.

3. Results and discussion

Listed in Table 2 are the 5-fold cross-validation results obtained by using the AAP antigenicity scale for different window sizes of 10, 12, 14, 16, 18 and 20. The parameter C is optimized over the range [10, 200] of the interval 10. As we can see from the table, the maximum prediction accuracy ranges from 64.39 to 68.07% with the window size of 20 achieving the largest value of 68.07%. Therefore, such a window size is used for further investigation.

We then compare the performance of linear SVM classifier using AAP antigenicity scale with that using AP scales. The AP scales include accessibility, flexibility, beta turn, antigenicity and hydrophilicity, which are thought to be correlated to the B-cell epitopes. Tables 3 and 4 show the results obtained by the 5-fold and jackknife cross-validation tests, respectively. From the two tables, we can see that the AAP antigenicity scale has outperformed the AP scales in all the cases.

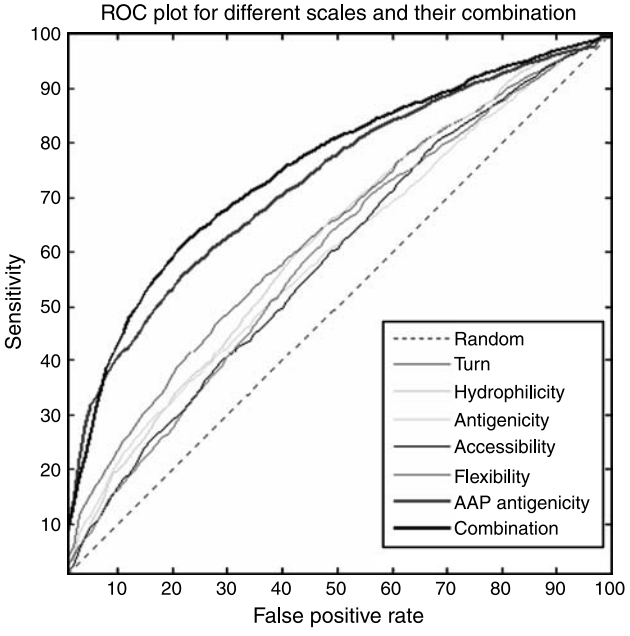

Fig. 2. ROC plot for AP scales, AAP antigenicity scale and their combination (for the colour version of this figure, the reader is referred to the online version of this paper under www.springerlink.com)

We also combined the AAP scale and the five AP scales using the SVM classifier in order to further improve the prediction accuracy. The last lines of Tables 3 and 4 show the results by the combination method. As we can see, the results obtained by such a combination are the best.

Finally, we also compared the performance of AAP antigenicity scale, AP scales, and their combination at different threshold levels with a ROC plot (Fig. 2). Again, as we can see from the plot, the combination method has the best performance at nearly all the threshold levels.

4. Conclusion

Prediction of the B-cell epitopes is an important but also very difficult and complicated problem. Based on the finding that B-cell epitopes favor particular AAPs, we have introduced a new scale, the so-called AAP antigenicity scale, to deal with the problem. The AAP antigenicity scale approach can give much better performance than various AP scale approaches used in the existing B-cell epitope prediction algorithms. The AAP antigenicity scale reflects the preference of B-cell epitopes for certain AAPs, which can be regarded as some kind of sequence-coupled effects. This accounts for its superiority over the AP scales. It is quite encouraging to see that if the AAP approach was combined with the existing AP approaches, the prediction quality would be further improved. It is expected that with the epitope data increasing, the performance of AAP antigenicity scale will be further improved.

References

Alix AJ (1999) Predictive estimation of protein linear epitopes by using the program PEOPLE. Vaccine 18: 311–314

Blythe MJ, Flower DR (2005) Benchmarking B cell epitope prediction: underperformance of existing methods. Protein Sci 14: 246–248

Cao Y, Liu S, Zhang L, Qin J, Wang J, Tang K (2006) Prediction of protein structural class with rough sets. BMC Bioinformatics 7: 20 [doi: 10.1186/1471-2105-7-20]

Chen C, Zhou X, Tian Y, Zou X, Cai P (2006) Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network. Anal Biochem 357: 116–121

Chou KC (1993) A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins. J Biol Chem 268: 16938–16948

Chou KC (1995) A sequence-coupled vector-projection model for predicting the specificity of GalNAc-transferase. Protein Sci 4: 1365–1383

Chou KC (1996) Review: prediction of HIV protease cleavage sites in proteins. Anal Biochem 233: 1–14

Chou KC (1997a) Prediction and classification of alpha-turn types. Biopolymers 42: 837–853

Chou KC (1997b) Prediction of beta-turns in proteins. J Peptide Res 49: 120–144

Chou KC (1999) Using pair-coupled amino acid composition to predict protein secondary structure content. J Protein Chem 18: 473–480

Chou KC (2000) Review: prediction of tight turns and their types in proteins. Anal Biochem 286: 1–16

Chou KC (2001a) Prediction of signal peptides using scaled window. Peptides 22: 1973–1979

Chou KC (2001b) Using subsite coupling to predict signal peptides. Protein Eng 14: 75–79

Chou KC (2002) Review: prediction of protein signal sequences. Curr Protein Peptide Sci 3: 615–622

Chou KC, Blinn JR (1997) Classification and prediction of beta-turn types. J Protein Chem 16: 575–595

Chou KC, Shen HB (2006a) Hum-PLoc: a novel ensemble classifier for predicting human protein subcellular localization. Biochem Biophys Res Commun 347: 150–157

Chou KC, Shen HB (2006b) Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers. J Proteome Res 5: 1888–1897

Chou KC, Zhang CT (1993) Studies on the specificity of HIV protease: an application of Markov chain theory. J Protein Chem 12: 709–724

Chou KC, Zhang CT (1995) Review: prediction of protein structural classes. Crit Rev Biochem Mol Biol 30: 275–349

Chou PY, Fasman GD (1978) Prediction of secondary structure of proteins from amino acid sequences. Adv Enzymol Rel Subjects Biochem 47: 45–148

Delacour H, Servonnet A, Perrot A, Vigezzi JF, Ramirez JM (2005) ROC (receiver operating characteristics) curve: principles and application in biology. Ann Biol Clin (Paris) 63: 145–154

Emini EA, Hughes JV, Perlow DS, Boger J (1985) Induction of hepatitis A virus-neutralizing antibody by a virus-specific synthetic peptide. J Virol 55: 836–839

Feng ZP (2001) Prediction of the subcellular location of prokaryotic proteins based on a new representation of the amino acid composition. Biopolymers 58: 491–499

Gao QB, Wang ZZ, Yan C, Du YH (2005a) Prediction of protein subcellular location using a combined feature of sequence. FEBS Lett 579: 3444–3448

Gao Y, Shao SH, Xiao X, Ding YS, Huang YS, Huang ZD, Chou KC (2005b) Using pseudo amino acid composition to predict protein subcellular location: approached with Lyapunov index, Bessel function, and Chebyshev filter. Amino Acids 28: 373–376

Guo YZ, Li M, Lu M, Wen Z, Wang K, Li G, Wu J (2006) Classifying G protein-coupled receptors and nuclear receptors based on protein power spectrum from fast Fourier transform. Amino Acids 30: 397–402

Karplus PA, Schulz GE (1985) Prediction of chain flexibility in proteins – a tool for the selection of peptide antigens. Naturwissenschaften 72: 212–213

Kolaskar AS, Tongaonkar PC (1990) A semi-empirical method for prediction of antigenic determinants on protein antigens. FEBS Lett 276: 172–174

Liu H, Yang J, Ling JG, Chou KC (2005) Prediction of protein signal sequences and their cleavage sites by statistical rulers. Biochem Biophys Res Commun 338: 1005–1011

Liu W, Chou KC (1999) Protein secondary structural content prediction. Protein Eng 12: 1041–1050

Odorico M, Pellequer JL (2003) BEPITOPE: predicting the location of continuous epitopes and patterns in proteins. J Mol Recogn 16: 20–22

Parker JM, Guo D, Hodges RS (1986) New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and X-ray-derived accessible sites. Biochemistry 25: 5425–5432

Saha S, Bhasin M, Raghava GP (2005) Bcipep: a database of B-cell epitopes. BMC Genomics 6: 79

Scholkopf B, Sung KK, Burges CJC, Girosi F, Niyogi P, Poggio T, Vapnik V (1997) Comparing support vector machines with Gaussian kernels to radial basis function classifiers. IEEE Trans Sign Proc 45: 2758–2765

Sollner J (2006) Selection and combination of machine learning classifiers for prediction of linear B-cell epitopes on proteins. J Mol Recogn 19: 209–214

Sollner J, Mayer B (2006) Machine learning approaches for prediction of linear B-cell epitopes on proteins. J Mol Recogn 19: 200–208

Sun XD, Huang RB (2006) Prediction of protein structural classes using support vector machines. Amino Acids 30: 469–475

Vapnik V (1998) Statistical learning theory. Wiley-Interscience, New York

Wen Z, Li M, Li Y, Guo Y, Wang K (2007) Delaunay triangulation with partial least squares projection to latent structures: a model for G-protein coupled receptors classification and fast structure recognition. Amino Acids (in press) (DOI: 10.1007/s00726-006-0341-y)

Xiao X, Shao S, Ding Y, Huang Z, Huang Y, Chou KC (2005) Using complexity measure factor to predict protein subcellular location. Amino Acids 28: 57–61

Zhang CT, Chou KC (1993) An alternate-subsite-coupled model for predicting HIV protease cleavage sites in proteins. Protein Eng 7: 65–73

Zhang SW, Pan Q, Zhang HC, Shao ZC, Shi JY (2006) Prediction protein homo-oligomer types by pseudo amino acid composition: approached with an improved feature extraction and naive Bayes feature fusion. Amino Acids 30: 461–468

Zhou GP (1998) An intriguing controversy over protein structural class prediction. J Protein Chem 17: 729–738

Zhou GP, Doctor K (2003) Subcellular location prediction of apoptosis proteins. Proteins Struct Funct Genet 50: 44–48

Authors' address: Kuo-Chen Chou, Gordon Life Science Institute, 13784 Torrey Del Mar Drive, San Diego, CA 92130, U.S.A.,
Fax: +1-858-484-1018; E-mail: kchou@san.rr.com