Nucleic Acids Research, 2005, Vol. 33, Web Server issue ***W1–W5***
doi:10.1093/nar/gki351

GPCRsclass: a web tool for the classification of amine type of G-protein-coupled receptors

Manoj Bhasin and G. P. S. Raghava*

Bioinformatics Center, Institute of Microbial Technology, Sector 39A, Chandigarh, India

Received November 15, 2004; Revised and Accepted December 6, 2004

ABSTRACT

The receptors of amine subfamily are specifically major drug targets for therapy of nervous disorders and psychiatric diseases. The recognition of novel amine type of receptors and their cognate ligands is of paramount interest for pharmaceutical companies. In the past, Chou and co-workers have shown that different types of amine receptors are correlated with their amino acid composition and are predictable on its basis with considerable accuracy [Elrod and Chou (2002) *Protein Eng.*, 15, 713–715]. This motivated us to develop a better method for the recognition of novel amine receptors and for their further classification. The method was developed on the basis of amino acid composition and dipeptide composition of proteins using support vector machine. The method was trained and tested on 167 proteins of amine subfamily of G-protein-coupled receptors (GPCRs). The method discriminated amine subfamily of GPCRs from globular proteins with Matthew's correlation coefficient of 0.98 and 0.99 using amino acid composition and dipeptide composition, respectively. In classifying different types of amine receptors using amino acid composition and dipeptide composition, the method achieved an accuracy of 89.8 and 96.4%, respectively. The performance of the method was evaluated using 5-fold cross-validation. The dipeptide composition based method predicted 67.6% of protein sequences with an accuracy of 100% with a reliability index ⩾5. A web server GPCRsclass has been developed for predicting amine-binding receptors from its amino acid sequence (http://www.imtech.res.in/raghava/gpcrsclass/).

INTRODUCTION

G-protein-coupled receptors (GPCRs) play a key role in cellular signaling pathways that regulate many basic physiological processes, such as neurotransmission, secretion, growth, cellular differentiation, inflammatory and immune responses (1). GPCRs consist of a single protein chain that crosses the membrane seven times (2). Currently known GPCRs include rhodopsin-like family, secretin-like receptor family, glutamate-like receptor, pheromone-like receptors, cAMP-like receptors and frizzled family of receptors (3). For sequences of >2000 GPCRs, the structure of only a single GPCR receptor (bovine rhodopsin) is available (4). There is a strong need for the detailed annotation of GPCRs from genomic data including recognition of GPCRs subfamilies and their types using computational tools. The rhodopsin-like receptors, which form a major superfamily of GPCRs, consist of ~60 families and subfamilies each one of which has many types of receptors (5). The typical strategies for identifying GPCRs and their types include similarity search based tools, such as BLAST, FASTA and motif finding tools (6). Although these tools are very successful in searching similar proteins, they fail when members of a subfamily are divergent in nature. To overcome this limitation, a number of tools based on composition and patterns of protein sequences have been developed. Most of the methods recognize the GPCRs and able to classify them up to major subfamilies. Recently, our group has also developed a method for recognizing and classifying GPCRs up to subfamily level (7). However, the method is not able to predict different types of receptors belonging to one subfamily.

The identification of novel type of GPCRs and their cognate ligands is the major focus of pharmaceutical companies. Hence, highly accurate identification of receptor types will solve the problem of efficacy and side effects of various drugs. Currently, few drugs are available that can bind to different types of receptors, hence have drastic side effects. Moreover, the mere understanding of different types of GPCRs and their substrate-binding properties will assist in finding novel drug target with minimum side effects. The experimental identifications of GPCR types are labor and cost-intensive task. The computational biology can provide a better alternative to develop a method for classifying different receptors of each. In the past, Elrod and Chou (8) showed that receptors of amine subfamily of rhodopsin-like superfamily have different amino acid compositions (8). They classified four types of amine-binding

AQ: Please check whether the editorial changes to this sentence retain the intended meaning.

AQ: Please check whether the editorial changes to this sentence retain the intended meaning.

*To whom the correspondence should be addressed. Tel: +91 172 2690557/2695225; Fax: +91 172 2690632/2690585; Email: raghava@imtech.res.in

AQ: Please check whether the format used for correspondence contact (telephone and fax numbers) is OK.

receptors (i) acetylcholine, (ii) adrenoceptor, (iii) dopamine and (iv) serotonin with overall accuracy of 83.23% using covariant discriminant analysis (8).

This motivated us to develop a highly accurate method for recognizing and classifying different types of amine receptors. The method has been developed using a two-step strategy. In first step, the method discriminates amine subfamily of GPCRs from other proteins, such as globular proteins. In second step, the method predicts the type of amine receptor using multiclass support vector machine (SVM). It has been shown in past that SVM is an elegant technique for the classification of biological data (9–16). The classification was achieved using amino acid and dipeptide composition. The method achieved a superior accuracy both in recognizing and classifying GPCRs. The results also proved the fact that dipeptide composition is a better feature for classifying the proteins. Dipeptide composition provides information about fraction of amino acids as well as local order, which is lacking in amino acid composition (17–18). To the best of authors' knowledge, there is no web server that allows recognition and classification of amine type of GPCRs. On the basis of the above study, an online web tool 'GPCRsclass' has been made available at http://www.imtech.res.in/raghava/gpcrsclass.

MATERIALS AND METHODS

Recognition of amine subfamily of GPCRs

At first step, the main aim is the recognition of novel GPCRs or discriminating GPCRs from the globular proteins. A SVM was trained to discriminate the GPCRs from other proteins. The training and testing was carried out on a dataset of 167 proteins of amine subfamily of GPCR. The dataset of 167 amine type of GPCRs was obtained from the study by Elrod and Chou (8). All the sequences of dataset were unique and complete (sequence fragments are removed from dataset). The training also required negative examples for discriminating GPCRs from other proteins. The dataset was extended by including 167 globular proteins obtained from the SCOP version 1.37 PDB90 domain database. The final dataset has equal number of positive and negative examples, so that the performance of the method can be evaluated using single parameter, such as accuracy.

Classification of amine subfamily of GPCRs

The amine subfamily of GPCR has major four types of receptors (acetylcholine, adrenoceptors, dopamine and serotonin). The dataset consisted of 167 sequences, of which 31 were acetylcholine, 44 adrenoceptors, 38 dopamine and 54 serotonin type of receptors. The classification of an unknown protein into one of the four types of amine receptors is a multiclass classification problem. In this regard, a series of binary classifiers were developed, which predict only a single type of amine receptors. Here, four SVMs were developed, one each for a particular type of amine receptor. The *i*th SVM was trained with all samples of the *i*th type receptors with positive label and samples of all other types of receptors as negative label. The SVMs trained in this way were referred as 1-v-r SVMs (7,19–20). In such classification, each of the unknown protein achieved four scores. An unknown protein was classified into the amine receptor type that corresponds to the 1-v-r SVM with highest output score.

SVM

The SVM was implemented using freely downloadable software SVM_light written by Joachims (21). SVMs nonlinearly map their *n*-dimensional input space into a high-dimensional feature space. In this high-dimensional feature space, a linear classifier is constructed and optimal hyperplane is constructed to separate the positive and negative examples. The SVM_light provides options for a number of inbuilt kernels, such as polynomial (given degree), radial basis function (RBF) and other regulatory parameters to achieve optimal classification of binary training and testing set. SVM requires the patterns of fixed length for testing and training. The proteins of variable length are transformed to fixed length format using amino acid and dipeptide composition (7,17,20,22). The amino acid composition provided the information of a protein in a vector of 20 dimensions. The dipeptide composition provides the information of protein in the form of a vector of 400 dimensions. The dipeptide composition encapsulates the information about fraction of amino acids as well as their local order.

Evaluation of performance

In order to evaluate the performance of prediction methods, jack-knife or limited cross-validations are the most commonly used procedures (7,17,19–20,23–28). During jack-knife cross-validation of N proteins, one protein is removed from the dataset, the training is performed on the remaining $N-1$ proteins and the testing is made on the removed protein (29). This process is repeated N times by removing each protein in turn. This cross-validation technique is time-consuming, so limited cross-validation is often performed when the dataset has larger number of proteins. In limited cross-validation, a set of proteins is divided into M equally balanced subsets. The method was trained or developed on $[(M-1)N]/M$ proteins and then tested on the remaining N/M proteins. This process is repeated M times, once for each subset. In this study, the performance of both amino acid and dipeptide composition based classifiers was evaluated through 5-fold cross-validation (17,19–20). The performance of the classifier developed at the first level (for recognizing proteins of amine subfamily) was evaluated using the standard threshold-dependent parameters, such as sensitivity, specificity, accuracy and Matthew's correlation coefficient (MCC). The performance of classifiers for classifying different types of amine receptors was evaluated by measuring accuracy and MCC as described by Hua and Sun (19).

Reliability index (RI)

The assignment of prediction reliability is important while using the machine learning techniques to predict types of amine receptors. RI was assigned on the basis of difference (Δ) between highest and second highest value of SVMs in multiclass classification. The RI for each sequence was defined by using Equation 1.

$$\mathrm{RI} = \begin{cases} \mathrm{INT}(\Delta * 5/3) + 1 & \text{if} \quad 0 \leqslant \Delta < 4 \\ 5 & \text{if} \quad \Delta \geqslant 4 \end{cases} \qquad \mathbf{1}$$

Table 1. The performance of amino acid composition and dipeptide composition based method in recognizing the GPCRs at different thresholds

Threshold	Amino acid composition				Dipeptide composition			
	Sen	Spe	Acc	MCC	Sen	Spe	Acc	MCC
−0.4	100	81.2	90.6	0.83	100	93.3	96.6	0.93
−0.2	100	86.7	93.2	0.88	99.4	99.4	99.4	0.98
0.0	99.4	96.9	98.2	0.96	99.4	100	99.7	0.99
0.2	97	99.4	98.2	0.96	82.0	100	91.0	0.83
0.4	95.2	100	97.6	0.95	52.1	100	76.1	0.59

Sen, sensitivity; Spe, specificity; Acc, accuracy.

Table 2. The performance of amino acid and dipeptide composition based method using different SVM kernels

Amine receptors	Amino acid composition based method				Dipeptide composition based method			
	REF kernel (γ = 500 and C = 3)		Polynomial kernel (d = 1 and C = 2000)		RBF kernel (γ = 100 and C = 10)		Polynomial kernel (d = 1 and C = 1000)	
	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC
Acteylcholine	87.1	0.92	90.3	0.90	93.6	0.96	93.6	0.96
Adrenoceptor	95.5	0.88	86.3	0.76	100	0.93	100	0.91
Dopamine	92.1	0.82	84.2	0.74	92.1	0.95	92.0	0.93
Serotonin	85.3	0.85	83.3	0.85	98.2	0.97	94.4	0.95
Overall	89.8	0.86	85.6	0.81	96.4	0.95	95.1	0.93

RESULTS AND DISCUSSION

The performance of the method in distinguishing amine subfamily of GPCRs from other globular proteins is shown in Table 1. The performance of the method is evaluated using a 5-fold cross-validation. The accuracy and MCC of amino acid composition based methods reached to 98.2% and 0.96, respectively, with RBF kernel at default threshold [0]. This demonstrates that amine subfamily of GPCRs can be well separated from other proteins on the basis of amino acid composition. The performance of SVM is better than covariant–discriminant analysis (83.3%) used by Elrod and Chou in (8). To further improve the accuracy, dipeptide composition was introduced instead of amino acid composition. The accuracy and MCC of the dipeptide composition based method are 99.7% and 0.99, respectively, using the RBF kernel (γ = 100 and C = 700). The performance of dipeptide composition based method is significantly better than the amino acid composition based classifier. The detailed performance of amino acid and dipeptide composition at different thresholds in term of sensitivity, specificity, accuracy and MCC is shown in Table 1. The results prove that amino acid composition as well as dipeptide composition can discriminate GPCRs from globular proteins with superior accuracy (>98%). The results are also consistent with our previous observation that dipeptide composition is better in classifying the proteins as compared with amino acid composition. The dipeptide composition is a better feature to encapsulate the global information about proteins as it provides information about fraction of amino acids as well as their local order.

Furthermore, to classify different types of amine receptors, a series of binary SVMs were constructed. The separate SVM modules have been developed for each type of receptors of amine subfamily. Each SVM is specific for one type of amine receptor. The overall accuracy of amino acid composition based classifier in classifying four types of receptors of amine subfamily is 89.8%. The classifiers were developed using polynomial kernels of various degree and RBF.

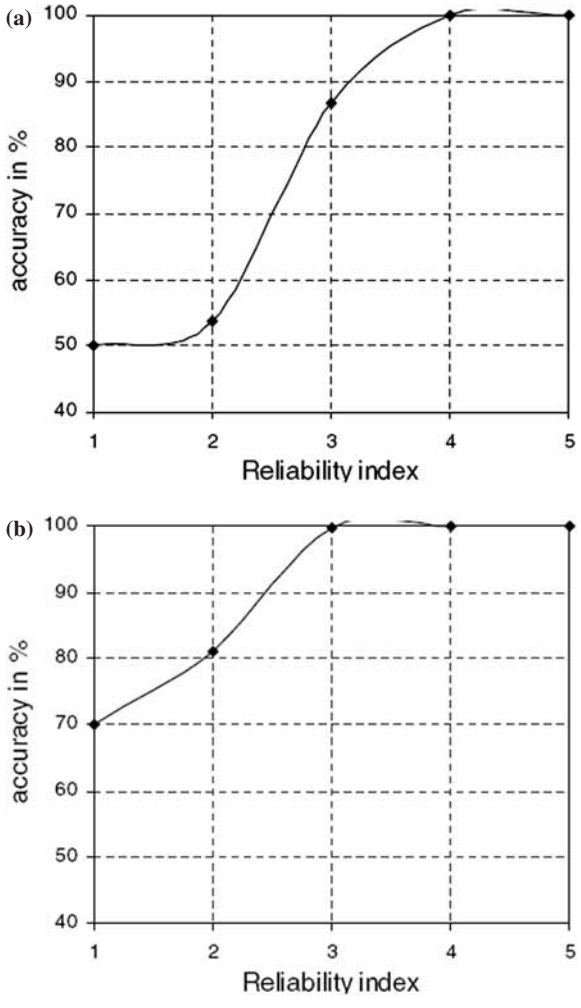

Figure 1. Expected accuracy of SVM classifier with an RI equal to a given value. The fraction of sequences that is predicted at a given RI is also shown on x-axis. (**a**) Amino acid composition. (**b**) Dipeptide composition.

AQ: Please check whether the editorial changes to this sentence retain the intended meaning.

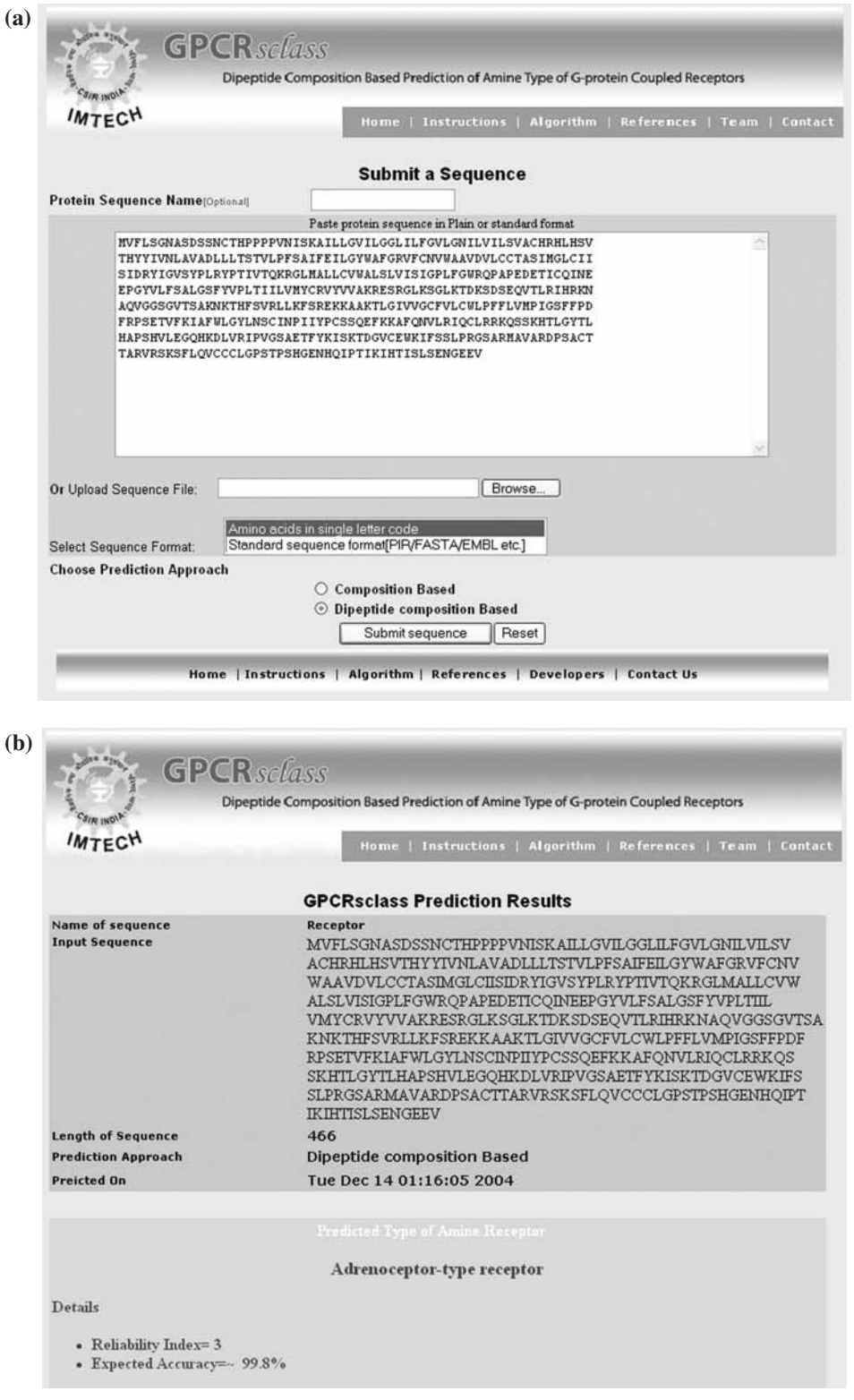

Figure 2. (**a**) Snapshot of input page of GPCRsclass server. (**b**) Snapshot of results obtained after the analysis of submission shown in Figure 1a.

The best results were achieved using RBF kernel with $\gamma = 500$ and $C = 3$. The detailed performance of amino acid composition based classifier for different amine receptors along with kernel parameters is shown in Table 2. For improving the accuracy, the classifier based on dipeptide composition of proteins was developed. The average accuracy and MCC of dipeptide composition based classifier are 96.4% and 0.95, respectively. The best results achieved using polynomial and RBF kernels along with kernel parameters for different types of amine receptors are shown in Table 2. As shown in

Table 2, the average accuracy of dipeptide composition based method was ~7% higher as compared with amino acid composition based classifier. The dipeptide composition based classifier classified four types of amine receptors with >92% accuracy. This proved that dipeptide composition is a better feature not only for recognizing but also for classifying different types of the amine receptors. This observation can also be extended to other types of receptors by establishing good training data.

Furthermore, to bring confidence in users about reliability of prediction, RI of amino acid composition as well as dipeptide composition based methods was measured. The RI provides information about the reliability or certainty of prediction. Figure 1a and b shows the expected accuracy of amino acid composition and dipeptide composition at different RI values. The expected accuracy of amino acid composition at RI = 5 is 100% with 62.2% of all sequences have RI = 5. Similarly for dipeptide composition at RI $\geqslant$ 3 the expected accuracy is 100% and about 74% of all sequences have RI $\geqslant$ 3.

These results suggest that types of GPCRs are predictable to a considerably accurate extent with amino acid composition as well as dipeptide composition. The development of such accurate and fast methods will speed up the identification of drug targets for curing various nervous system diseases.

Description of server

GPCRsclass is freely available at http://www.imtech.res.in/raghava/gpcrsclass/. The common gateway interface script of GPCRsclass is written using PERL version 5.03. GPCRsclass server is installed on a Sun Server (420E) under UNIX (Solaris 7) environment. The user can provide the input sequence by cut-paste or directly uploading sequence file from disk. The server accepts the sequence in raw format as well as in standard format, such as EMBL, FASTA and GCG acceptable to ReadSeq (developed by Dr Don Gilbert). A snapshot sequence page of server is shown in Figure 2a. User can predict the type of amine receptors based on either amino acid composition or dipeptide composition. On submission the server will give results in user-friendly format as shown in Figure 2b.The prediction results also provide information about prediction reliability (RI) and expected accuracy.

ACKNOWLEDGEMENTS

AQ: Please provide details for 'xxxx' and complete the sentence.

Funding to pay the Open Access publication charges for this article was provided by xxxxx.

REFERENCES

AQ: References (1–8) have been renumbered to meet the journal requirement that first citations be in numerical order. Please check that each citation number points to the correct reference.

1. Attwood,T.K., Croning,M.D. and Gaulton,A. (2002) Deriving structural and functional insights from a ligand-based hierarchical classification of G protein-coupled receptors. *Protein Eng*., **15**, 7–12.
2. Horn,F., Weare,J., Beukers,M.W., Hörsch,S., Bairoch,A., Chen,W., Edvardsen,Ø., Campagne,F. and Vriend,G. (1998) GPCRDB: an information system for G protein-coupled receptors. *Nucleic Acids Res*., **26**, 277–281.
3. Sadowski,M.I. and Parish,J.H. (2003) Automated generation and refinement of protein signatures: case study with G-protein coupled receptors. *Bioinformatics*, **19**, 727–734.
4. Karchin,R., Karplus,K. and Haussler,D. (2002) Classifying G-protein coupled receptors with support vector machines. *Bioinformatics*, **18**, 147–159.
5. Horn,F., Vriend,G. and Cohen,F.E. (2001) Collecting and harvesting biological data: the GPCRDB and NucleaRDB information systems. *Nucleic Acids Res*., **29**, 346–349.
6. Lapinsh,M., Gutcaits,A., Prusis,P., Post,C., Lundstedt,T. and Wikberg,J.E. (2002) Classification of G-protein coupled receptors by alignment-independent extraction of principal chemical properties of primary amino acid sequences. *Protein Sci*., **11**, 795–805.
7. Bhasin,M. and Raghava,G.P.S. (2004) GPCRpred: an SVM-based method for prediction of families and subfamilies of G-protein coupled receptors. *Nucleic Acids Res*., **32**, W383–W389.
8. Elrod,D.W. and Chou,K.C. (2002) A study on the correlation of G-protein-coupled receptor types with amino acid composition. *Protein Eng*., **15**, 713–715.
9. Bhasin,M. and Raghava,G.P. (2004) SVM based method for predicting HLA-DRB1*0401 binding peptides in an antigen sequence. *Bioinformatics*, **20**, 421–423.
10. Chou,K.C. and Cai,Y.D. (2002) Using functional domain composition and support vector machines for prediction of protein subcellular location. *J. Biol. Chem*., **277**, 45765–45769.
11. Cai,Y.D., Liu,X.J., Xu,X.B. and Chou,K.C. (2002) Support vector machines for prediction of protein subcellular location by incorporating quasi-sequence-order effect. *J. Cell. Biochem*., **84**, 343–348.
12. Cai,Y.D., Pong-Wong,R., Feng,K., Jen,J.C.H. and Chou,K.C. (2004) Application of SVM to predict membrane protein types. *J. Theor. Biol*., **226**, 373–376.
13. Cai,Y.D., Zhou,G.P. and Chou,K.C. (2003) Support vector machines for predicting membrane protein types by using functional domain composition. *Biophys. J*., **84**, 3257–3263.
14. Cai,Y.D., Zhou,G.P., Jen,C.H., Lin,S.L. and Chou,K.C. (2004) Identify catalytic triads of serine hydrolases by support vector machines. *J. Theor. Biol*., **228**, 551–557.
15. Wang,M., Yang,J., Liu,G.P., Xu,Z.J. and Chou,K.C. (2004) Weighted-support vector machines for predicting membrane protein types based on pseudo amino acid composition. *Protein Eng. Des. Sel*., **17**, 509–516.
16. Yang,Z.R. and Chou,K.C. (2004) Bio-support vector machines for computational proteomics. *Bioinformatics*, **20**, 735–741.
17. Bhasin,M. and Raghava,G.P.S. (2004) ELSPred:SVM based prediction of subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. *Nucleic Acids Res*., **32**, W414–W419.
18. Chou,K.C. (1999) A key driving force in determination of protein structural classes. *Biochem. Biophys. Res. Commun*., **264**, 216–224.
19. Hua,S. and Sun,Z. (2001) Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, **17**, 721–728.
20. Bhasin,M. and Raghava,G.P.S. (2004) Classification of nuclear receptors based on amino acid composition and dipeptide composition. *J. Biol. Chem*., **279**, 23262–23266.
21. Joachims,T. (1999) Making large-Scale SVM Learning Practical. In Scholkopf,B., Burges,C. and Smola,A. (eds.), *Advances in Kernel Methods Support Vector Learning*. MIT Press, Cambridge, Massachusetts, London, England.
22. Liu,W. and Chou,K.C. (1999) Protein secondary structural content prediction. *Protein Eng*., **12**, 1041–1050.
23. Feng,Z.P. (2001) Prediction of the subcellular location of prokaryotic proteins based on a new representation of the amino acid composition. *Biopolymers*, **58**, 491–499.
24. Pan,Y.X., Zhang,Z.Z., Guo,Z.M., Feng,G.Y., Huang,Z.D. and He,L. (2003) Application of pseudo amino acid composition for predicting protein subcellular location: stochastic signal processing approach. *J. Protein Chem*., **22**, 395–402.
25. Yuan,Z. (1999) Prediction of protein subcellular locations using Markov chain models. *FEBS Lett*., **451**, 23–26.
26. Zhou,G.P. (1998) An intriguing controversy over protein structural class prediction. *J. Protein Chem*., **17**, 729–738.
27. Zhou,G.P. and Assa-Munt,N. (2001) Some insights into protein structural class prediction. *Proteins*, **44**, 57–59.
28. Zhou,G.P. and Doctor,K. (2003) Subcellular location prediction of apoptosis proteins. *Proteins*, **50**, 44–48.
29. Chou,K.C. and Zhang,C.T. (1995) Review: Prediction of protein structural classes. *Crit. Rev. Biochem. Mol. Biol*., **30**, 275–349.