W414–W419 *Nucleic Acids Research, 2004, Vol. 32, Web Server issue*
DOI: 10.1093/nar/gkh350

ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST

Manoj Bhasin and G. P. S. Raghava*

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

Received January 14, 2004; Accepted February 9, 2004

ABSTRACT

Automated prediction of subcellular localization of proteins is an important step in the functional annotation of genomes. The existing subcellular localization prediction methods are based on either amino acid composition or N-terminal characteristics of the proteins. In this paper, support vector machine (SVM) has been used to predict the subcellular location of eukaryotic proteins from their different features such as amino acid composition, dipeptide composition and physico-chemical properties. The SVM module based on dipeptide composition performed better than the SVM modules based on amino acid composition or physico-chemical properties. In addition, PSI-BLAST was also used to search the query sequence against the dataset of proteins (experimentally annotated proteins) to predict its subcellular location. In order to improve the prediction accuracy, we developed a hybrid module using all features of a protein, which consisted of an input vector of 458 dimensions (400 dipeptide compositions, 33 properties, 20 amino acid compositions of the protein and 5 from PSI-BLAST output). Using this hybrid approach, the prediction accuracies of nuclear, cytoplasmic, mitochondrial and extracellular proteins reached 95.3, 85.2, 68.2 and 88.9%, respectively. The overall prediction accuracy of SVM modules based on amino acid composition, physico-chemical properties, dipeptide composition and the hybrid approach was 78.1, 77.8, 82.9 and 88.0%, respectively. The accuracy of all the modules was evaluated using a 5-fold cross-validation technique. Assigning a reliability index (reliability index $\geqslant$3), 73.5% of prediction can be made with an accuracy of 96.4%. Based on the above approach, an online web server ESLpred was developed, which is available at http://www.imtech.res.in/raghava/eslpred/.

INTRODUCTION

Large-scale genome sequencing projects make interpretation of genomic sequence data increasingly important, so does the need to functionally annotate this data. The determination of subcellular localization of a protein can provide important clues to elucidate the function of the protein (1). Therefore, prediction of subcellular localization of proteins is an important step in understanding the biochemical function of proteins. In the past, various methods have been developed to predict the subcellular location of proteins using different approaches (2). The similarity search in which a sequence is searched against an experimentally annotated database, is a technique commonly used to assign function to a protein, including its subcellular location (3). This approach fails in the absence of significant similarity between query and target protein sequences (3). Another way to predict subcellular localization of proteins is to identify sequence motifs such as signal peptide or nuclear localization signal (4). The major limitation of motif-based methods is that all proteins residing in a compartment do not have universal motifs.

To overcome these limitations, in the past numerous studies have been carried out to predict subcellular localization based on the features of protein sequence. The subcellular localization prediction methods are based either on recognition of N-terminal sorting signals or on the composition of amino acids. Since 1991, numerous methods have been developed to predict subcellular localization and are available online (PSORT I (5) and PSORT-B (6) for prokaryotic organisms, iPSORT (7) and TargetP (8) for eukaryotes, and SubLoc (2) and NNPSL (3) for both prokaryotes and eukaryotes). NNPSL and SubLoc were developed using artificial neural network (ANN) and support vector machine (SVM), respectively, on the basis of composition of amino acids. The accuracies of these methods vary remarkably

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

from each other though they are developed on the same dataset of proteins. The prediction accuracy for eukaryotic proteins was less than for prokaryotic proteins due to the complex organization of eukaryotes (9). Hence, there is a need to improve prediction accuracy of subcellular localization of eukaryotic proteins.

In this paper, a systematic attempt has been made to achieve higher prediction accuracy for subcellular localization of eukaryotic proteins from their different features. The SVM modules were developed based on the following features of a protein: (i) amino acid composition (commonly used in the literature for classification of proteins), (ii) overall physico-chemical properties (e.g. hydrophobicity, hydrophilicity, polarity) and (iii) dipeptide compositions (e.g. ala–ala, ala–leu, val–ser). The prediction accuracy of the dipeptide-based SVM module is superior to the amino acid composition and physico-chemical properties based modules. In addition, a similarity search based module, EuPSI-BLAST, was also constructed using PSI-BLAST to predict the localization of a protein. Finally, a hybrid SVM module was developed using all three features of proteins mentioned above and prediction results of EuPSI-BLAST. Development of the hybrid module resulted in significant improvement in prediction accuracy. The development of the hybrid module using this novel approach fulfilled the goal behind the development of a more reliable method. This method can complement the existing subcellular localization prediction methods and can assist in the development of automated genome annotation tools.

MATERIALS AND METHODS

The dataset used in the present work was obtained from http://www.doe-mbi.ucla.edu/%7Eastrid/astrid.html, which was also used in the development of SubLoc (2) and NNPSL (3). This dataset was generated from version 33.0 of SWISS-PROT (10) by Reinhardt and Hubbard. The dataset consisted of complete and non-redundant proteins with less than 90% sequence identity whose subcellular localization is experimentally determined. This dataset consisted of a total of 2427 eukaryotic proteins (1097 nuclear, 684 cytoplasmic, 321 mitochondrial and 325 extracellular proteins).

Evaluation of ESLpred

The performance modules constructed in this study were evaluated using a 5-fold cross-validation technique. In the 5-fold cross-validation, the relevant dataset was partitioned randomly into five equally sized sets. The training and testing was carried out five times, each time using one distinct set for testing and the remaining four sets for training. For evaluating the performance of various modules, accuracy and Matthew's correlation coefficient (MCC) were calculated using the following equations:

$$\text{Accuracy}(x) = \frac{p(x)}{\text{Exp}(x)}, \quad \mathbf{1}$$

$$\text{MCC}(x) = \frac{p(x)n(x) - u(x)o(x)}{\sqrt{[p(x) + u(x)][p(x) + o(x)][n(x) + u(x)][n(x) + o(x)]}}, \quad \mathbf{2}$$

where x can be any subcellular location (nuclear, cytoplasm, extracellular and mitochondria), $\exp(x)$ is the number of sequences observed in location x, $p(x)$ is the number of correctly predicted sequences of location x, $n(x)$ is the number of correctly predicted sequences not of location x, $u(x)$ is the number of under-predicted sequences and $o(x)$ is the number of over-predicted sequences.

Support vector machine

SVMs are universal approximators based on statistical and optimization theory. The SVM is particularly attractive to biological sequence analysis due to its ability to handle noise, large dataset and large input spaces (11). Further details about the SVM can be obtained from Vapnik's papers (12) or http://www.imtech.res.in/raghava/eslpred/algo.html. In the present study, we have used SVM_light to predict the subcellular localization of proteins. This software is freely downloadable from http://www.cs.cornell.edu/People/tj/svm_light/. The software enables the users to define a number of parameters and also allows a choice of inbuilt kernel function, including linear, RBF and Polynomial. The parameters except kernel functions and regulatory parameters C were kept constant during the training. The prediction of subcellular localization is a multi-class classification problem. We developed a series of binary classifiers to handle the multi-classification problem. We constructed N SVMs for N-class classification. Here, the class number was equal to four for eukaryotic sequences. The ith SVM was trained with all samples in the ith class with positive labels and all other samples with negative labels. In this way, four SVMs were constructed for subcellular localization of protein to nuclear, cytoplasm, extracellular and mitochondria. An unknown sample was classified into the class that corresponded to the SVM with highest output score.

Protein features

Amino acid composition. Amino acid composition is the fraction of each amino acid in a protein. The fraction of all 20 natural amino acids was calculated using the following equation:

$$\text{fraction of amino acid } i = \frac{\text{total number of amino acid } i}{\text{total number of amino acids in protein}}, \quad \mathbf{3}$$

where i can be any amino acid.

Composition of physico-chemical properties. The 33 physico-chemical properties were used to represent proteins as shown in Table S1 of the supplementary material (13). The values of each physio-chemical property for all 20 amino acids were normalized between 0 and 1 using the standard conversion formula. The input vector has 33 scalar values, each representing the average value of a distinct physico-chemical property of protein.

Dipeptide composition. Dipeptide composition was used to encapsulate the global information about each protein sequence, which gives a fixed pattern length of 400 (20×20). This representation encompassed the information about amino acid composition along local order of amino acid. The fraction

of each dipeptide was calculated using following equation:

$$\text{fraction of dep}(i) = \frac{\text{total number of dep}(i)}{\text{total number all possible dipeptides}}, \quad \mathbf{4}$$

where dep(i) is one out of 400 dipeptides.

EuPSI-BLAST

A module EuPSI-BLAST was designed in which query sequence was searched against a database of 2427 eukaryotic proteins using PSI-BLAST. Three iterations of PSI-BLAST were carried out at a cut-off E-value of 0.001. PSI-BLAST was used instead of normal standard BLAST because PSI-BLAST has the capability to detect remote homologies. The module could predict any of the four localizations (cytoplasmic, nuclear, mitochondrial or extracellular) depending upon the similarity of the query protein to the proteins in the dataset. The module would return 'unknown subcellular localization' if no significant similarity was obtained.

Input for hybrid SVM module

This module uses complete information about protein, i.e. amino acid composition, dipeptide composition, physico-chemical properties and PSI-BLAST output. The overall architecture of the hybrid module is shown at http://www.imtech.res.in/raghava/eslpred/algo.html. SVM was provided with an input vector of 458 dimensions that consisted of 20 for amino acid composition, 400 for dipeptide composition, 33 for physiochemical properties and five for PSI-BLAST output. The BLAST output was converted to binary variables using the following representations:

$$\begin{pmatrix} \text{Nuclear} & \rightarrow & 1 & 0 & 0 & 0 & 0 \\ \text{Cytoplasmic} & \rightarrow & 0 & 1 & 0 & 0 & 0 \\ \text{Mitochondrial} & \rightarrow & 0 & 0 & 1 & 0 & 0 \\ \text{Extracellular} & \rightarrow & 0 & 0 & 0 & 1 & 0 \\ \text{Unknown} & \rightarrow & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Reliability index

The reliability index (RI) is a commonly used measure of prediction which provides confidence about a prediction to the users. The RI assignment is a useful indication of the level of certainty in the prediction for a particular sequence. We have followed this simple strategy for assigning the RI as used in the past (2,3). The RI was assigned according to the difference (Δ) between the highest and second highest SVM output scores. We have also computed the reliability score of our prediction method based on the hybrid approach using the following equation:

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

PREDICTION RESULTS

The performance of all the modules developed in this study is shown in Table 1. The performance of all modules was evaluated through 5-fold cross-validation. The composition-based SVM module (kernel = RBF, $\gamma = 16$ and $C = 1000$) was able to predict with 78.1% accuracy. The performance of this composition-based module was nearly equal to SubLoc (2). The physico-chemical properties-based SVM module predicted subcellular localization of protein with slightly lower accuracy (77.8%) in comparison with the amino acid composition-based module. In the case of the physio-chemical properties-based module the best results were achieved with the RBF kernel ($\gamma = 15$ and $C = 1000$). These results indicated that 33 physico-chemical properties could predict the subcellular localization of a protein with fair accuracy. In the case of the dipeptides composition-based module the performance of the RBF kernel ($\gamma = 200$ and $C = 1000$) was nearly 5% better than the amino acid composition- and properties-based SVM modules (Table 1). Thus, dipeptide composition, which provided information about amino acid composition as well as local order of amino acids, is a better feature for predicting subcellular localization of eukaryotic proteins.

The results of the EuPSI-BLAST module were also evaluated through 5-fold cross-validation. The module predicted nuclear, cytoplasmic, mitochondrial and extracellular proteins with 84.5, 77.6, 54.8 and 86.7% accuracy, respectively. Beside this, a module based on standard BLAST (14) was also constructed. The performance of this module was poor (Table 1). In the case of standard BLAST during cross-validation, no significant hit was obtained for 508 proteins out of 2427 proteins, whereas it was only 362 proteins for which no significant hit was found in the case of PSI-BLAST. This observation strengthened the fact the PSI-BLAST is able to detect the protein having remote homology.

Table 1. The performance of various modules including SVM modules based on various features of protein sequence, standard BLAST and PSI-BLAST

Approach	Nuclear		Cytoplasmic		Mitochondrial		Extracellular		Overall
	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC
Composition-based (A)	86.1	0.73	76.9	0.64	55.5	0.54	76.0	0.76	**78.1**
Properties Based (B)	85.6	0.73	74.6	0.64	59.2	0.55	76.6	0.74	**77.8**
Dipeptide Based (C)	92.7	0.79	80.2	0.71	58.8	0.62	79.0	0.83	**82.9**
EuPSI-BLAST (C)	84.5	—	77.6	—	54.8	—	86.7	—	—
EuBLAST	76.5	—	78.0	—	57.0	—	82.7	—	—
Hybrid1 (B + C)	93.3	0.81	81.1	0.74	64.5	0.67	82.4	0.85	**84.6**
Hybrid2 (A + B + C)	93.2	0.81	80.6	0.73	65.1	0.67	83.4	0.86	**84.6**
Hybrid (A + B + C + D)	95.3	0.87	85.2	0.79	68.2	0.69	88.9	0.91	**88.0**

ACC: Accuracy; MCC: Matthew's correlation coefficient.

To further improve the prediction accuracy, hybrid modules on the basis of various features of proteins were constructed. The first hybrid module (hybrid1) was developed on the basis of the dipeptide composition and physico-chemical properties of proteins. The prediction accuracy of the hybrid1 module was 84.6%, which was better than any individual feature-based module. Another module (hybrid2) was developed on the basis of amino acid composition, dipeptide composition and physico-chemical properties; its performance was similar to the hybrid1 module. Finally, a hybrid module based on all features of proteins and PSI-BLAST information was developed. This hybrid module used an input vector of 458 dimensions, comprising 20 for amino acid compositions, 400 for dipeptide composition, 33 for various physico-chemical properties and 5 for PSI-BLAST output. As shown in Table 1, the performance of this hybrid module is better than any individual feature-based or other hybrid modules (hybrid1 and hybrid2). The detailed performance of the hybrid module in terms of accuracy and MCC is shown in Table 1. Finally, a hybrid module with the RBF kernel ($\gamma = 50$ and $C = 1000$), which used all features of proteins and EuPSI-BLAST information, was able to achieve 88.0% accuracy.

The RI assignment was also carried out to know the prediction reliability. The prediction accuracy with RI equal to a given value was calculated as shown in the supplementary data (http://www.imtech.res.in/raghava/eslpred/algo.html). The RI curve depicted that the expected accuracy of sequences with RI = 3 was 94.4%, which is better than existing methods [e.g. SubLoc (2) and NNPSL (3)]. Another calculation showed that nearly 74% of sequences have RI $\geq$ 3, and the expected accuracy of these sequences was 96.4%.

Comparison with other prediction methods

The performance of the hybrid module developed in this study was compared with existing methods such as Subloc and NNPSL (2,3), which were also developed from the same dataset. The results demonstrated that overall prediction accuracy of a hybrid module is nearly 10% greater than the composition-based method SubLoc (2). The MCC of the hybrid module for each subcellular location was higher than the corresponding one for the SubLoc method. The prediction accuracy of the hybrid module was >20% higher than that of the neural network-based method NNPSL (3).

Description of the server

All the modules constructed in this study have been implemented on the World Wide Web as a dynamic web server 'ESLpred', which is available at http://www.imtech.res.in/raghava/eslpred. All the CGI scripts of the method were written in PERL5.0 and the interface was designed using HTML. The server runs on SUN server 420R under the Solaris environment. The SVM and PSI-BLAST were implemented by obtaining SVM_light from http://www.cs.cornell.edu/People/tj/svm_light/ and PSI-BLAST from http://www.ncbi.nlm.nih.gov/blast/. It is a user-friendly web server and allows users to submit their protein sequence in one of the standard formats such as FASTA, GenBank, EMBL, GCG or plain format (Figure 1). The user can input their sequence by typing or pasting in box or by using the file upload facility. The server uses the ReadSeq program to read the input sequence. The server provides an option to select the prediction approach. In the case of default prediction, the server

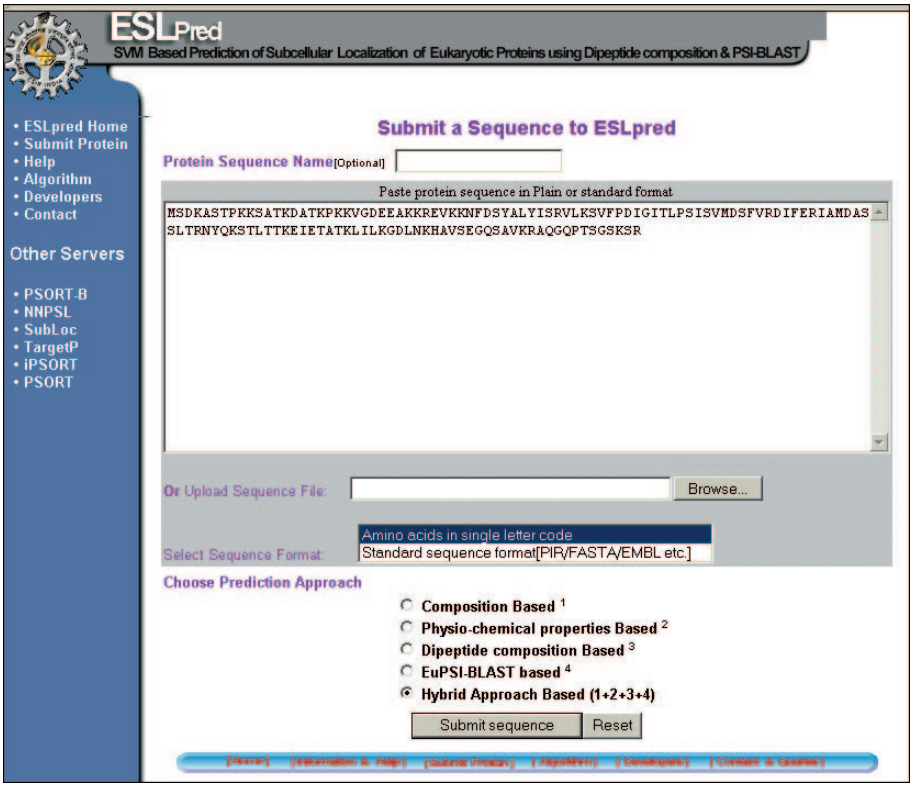

Figure 1. A snapshot of the query submission page of ESLpred server.

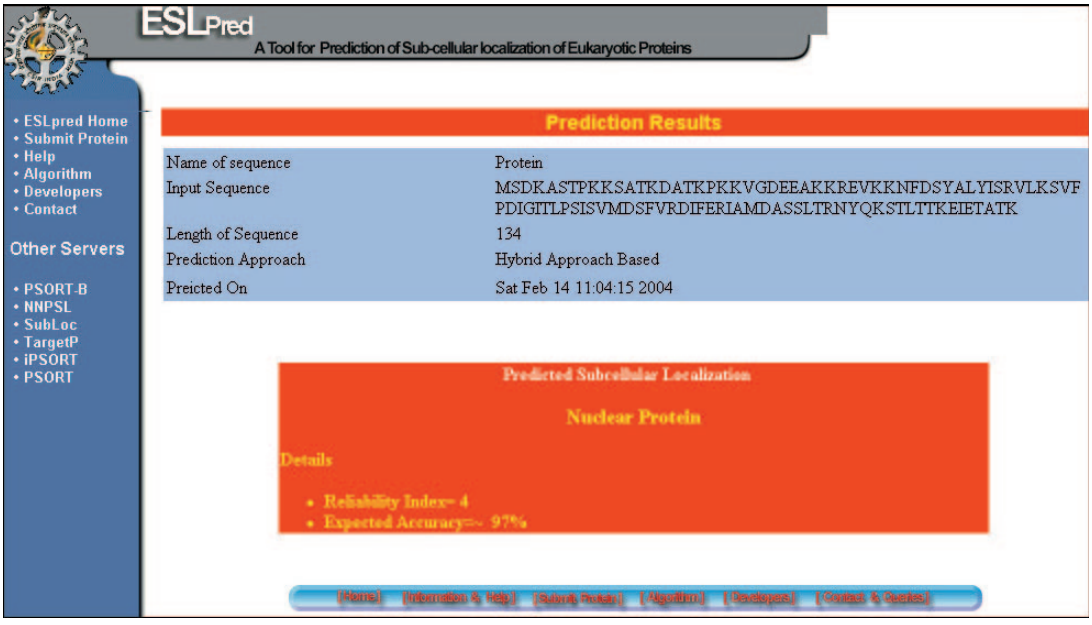

Figure 2. Prediction results of a query protein using the SVM-based hybrid module.

uses the hybrid module for prediction. The server presents the results of comprehensive analysis in user-friendly format (Figure 2).

DISCUSSION

In general, artificial intelligence (AI) based techniques such as SVMs and neural networks are elegant approaches for the extraction of complex patterns from biological sequence data. These techniques are highly successful for residue state prediction where fixed window/pattern length is used (15). The major limitation of the AI techniques is that they need patterns/input units of fixed length. This is the major reason for the failure of the AI techniques in the classification of proteins (e.g. subcellular localization prediction, fold recognition) because similar/homologous proteins often have variable length. In order to overcome this problem, a fixed-length pattern must be generated for proteins, for AI techniques to be implemented.

The percentage composition of amino acids, which gives a fixed pattern length of 20, is commonly used by AI techniques for the classification of proteins. This strategy has been used previously for developing the method for subcellular localization prediction of eukaryotic and prokaryotic proteins (2,3). However, this approach provides information only about the amino acid frequency, but no information about the local order of amino acids (16). To provide the information about frequency and local order of amino acids , dipeptide composition (instead of amino acid composition) can be used as the input unit to AI techniques. Dipeptide composition gives a fixed pattern length of 400. Dipeptide composition is widely used in the development of methods for fold prediction (17). The prediction accuracy of the dipeptide composition-based method should be higher than that of amino acid composition based methods (18). More information about the protein sequence can be encapsulated using tripeptide composition. Tripeptide composition gives a fixed pattern length of 8000, which is commonly used in similarity searching in BLAST and FASTA (14,19). In the case of tripeptide composition, ANN and SVM are unable to handle the noise due to the large number of input units and number of missing tripeptides in a protein. Therefore, in this paper, we have constructed a SVM module on the basis of the dipeptide composition of a protein. This module is able to predict the subcellular location of a protein with overall accuracy of 82.9%, as shown in Table 1.

The physico-chemical properties of a protein are yet another alternative way to provide the global information of a protein in the form of fixed pattern length. In this paper, a module using 33 physico-chemical properties of proteins was developed to encapsulate the global information of a protein. A fixed pattern length of 33 was used, where each unit corresponded to a property of a protein. The SVM module based on this approach was able to predict the subcellular localization of proteins with fair accuracy (77.8%), as shown in Table 1.

To further improve prediction accuracy, we have devised methodologies to encapsulate more comprehensive information of a protein. A SVM-based module (hybrid) was constructed on the basis of comprehensive information about proteins including amino acid composition, physico-chemical properties, dipeptide composition and PSI-BLAST results. The hybrid module predicted the subcellular localization of a protein more accurately than the rest of the modules developed in this study. These results confirmed that our approach is capable of capturing more information about a protein that is crucial for detecting subcellular localization of proteins. Thus, providing more comprehensive information can be useful in enhancing the prediction accuracy of fold or tertiary structure prediction methods.

In conclusion, a new method for subcellular localization of a eukaryotic protein is presented. This method will nicely complement the existing subcellular localization prediction

methods. It will assist in assigning the subcellular location or function of proteins more reliably. The authors believe that the prediction method presented here would be useful for the annotation of the piled-up genomic data.

ACKNOWLEDGEMENTS

The authors are thankful to the Council of Scientific and Industrial Research (CSIR) and Department of Biotechnology (DBT), Government of India for financial assistance. This report has IMTECH communication no. 04/2004.

REFERENCES

1. Eisenhaber,F. and Bork,P. (1998) Wanted: subcellular localization of proteins based on sequence. *Trends Cell Biol*., **8**, 69–70.
2. Hua,S. and Sun,Z. (2001) Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, **17**, 721–728.
3. Reinhardt,A. and Hubbard,T. (1998) Using neural networks for prediction of the subcellular location of proteins. *Nucleic Acids Res*., **26**, 2230–2236.
4. Fujiwara,Y. and Asogawa,M. (2001) Prediction of subcellular localizations using amino acid composition and order. *Genome Inform. Ser. Workshop Genome Inform*., **12**, 103–112.
5. Nakai,K. and Kanehisa,M. (1991) Expert system for predicting protein localization sites in Gram-negative bacteria. *Proteins*, **11**, 95–110.
6. Gardy,J.L., Spencer,C., Wang,K., Ester,M., Tusnady,G.E., Simon,I., Hua,S., deFays,K., Lambert,C., Nakai,K. *et al*. (2003) PSORT-B: Improving protein subcellular localization prediction for Gram-negative bacteria. *Nucleic Acids Res*., **31**, 3613–3617.
7. Bannai,H., Tamada,Y., Maruyama,O., Nakai,K. and Miyano,S. (2002) Extensive feature detection of N-terminal protein sorting signals. *Bioinformatics*, **18**, 298–305.
8. Emanuelsson,O., Nielsen,H., Brunak,S. and von Heijne,G. (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol*., **300**, 1005–1016.
9. Nakai,K. and Horton,P. (1999) PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem. Sci*., **24**, 34–36.
10. Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequences database and its supplement TrEMBL in 2000. *Nucleic Acids Res*., **28**, 45–48.
11. Zavaljevski,N., Stevens,F.J. and Reifman,J. (2002) Support vector machines with selective kernel scaling for protein classification and identification of key amino acid positions. *Bioinformatics*, **18**, 689–696.
12. 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, MA, London, England .
13. Bhasin,M. and Raghava,G.P.S. (2004) Analysis and prediction of quantitative affinity of TAP binding peptides using cascade SVM. *Protein Sci*., **13**, 596–607.
14. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J., (1990). Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.
15. Krogh,A. and Riis,S.K. (1996) Prediction of β sheets in protein. In Touretzky,D.S., Mozer,M.C., Hasaselmo,M.E. (eds), *Advances in Neural Information Processing System 8*. MIT Press, Cambridge, MA, pp. 917–923.
16. Shepherd,A.J., Gorse,D. and Thornton,J.M. (2003) A novel approach to the recognition of protein architecture from sequence using Fourier analysis and neural networks. *Proteins*, **50**, 290–302.
17. Reczko,M. and Bohr,H. (1995) The DEF database of sequence based protein fold class prediction. *Nucleic Acid Res*., **22**, 3616–3619.
18. Grassmann,J., Reczko,M., Suhai,S. and Edler,L. (1999) Protein fold class prediction: new methods of statistical classification. In Lengauer,T., Schneider,R., Bork,P., Brutlag,D.L., Glasgow,J.I., Mewes,H.-W. and Zimmer,R. (eds), *Proceedings of Seventh International Conference on Intelligent System for Molecular Biology* (ISMB'99), AAAI, Heidelberg, Germany, pp. 106–112.
19. Pearson,W.R. and Lipman,D.J. (1988). Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci., USA*, **85**, 2444–2448.