BMC Bioinformatics

BioMed Central

Research article

Open Access

Esub8: A novel tool to predict protein subcellular localizations in eukaryotic organisms

Qinghua Cui, Tianzi Jiang*, Bing Liu and Songde Ma

Address: National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100080, P. R. China

Email: Qinghua Cui - qhcui@nlpr.ia.ac.cn; Tianzi Jiang* - jiangtz@nlpr.ia.ac.cn; Bing Liu - bliu@nlpr.ia.ac.cn; Songde Ma - masd@nlpr.ia.ac.cn
* Corresponding author

Published: 27 May 2004

BMC Bioinformatics 2004, **5**:66

Received: 16 December 2003
Accepted: 27 May 2004

This article is available from: http://www.biomedcentral.com/1471-2105/5/66

Abstract

Background: Subcellular localization of a new protein sequence is very important and fruitful for understanding its function. As the number of new genomes has dramatically increased over recent years, a reliable and efficient system to predict protein subcellular location is urgently needed.

Results: Esub8 was developed to predict protein subcellular localizations for eukaryotic proteins based on amino acid composition. In this research, the proteins are classified into the following eight groups: chloroplast, cytoplasm, extracellular, Golgi apparatus, lysosome, mitochondria, nucleus and peroxisome. We know subcellular localization is a typical classification problem; consequently, a one-against-one (1-v-1) multi-class support vector machine was introduced to construct the classifier. Unlike previous methods, ours considers the order information of protein sequences by a different method. Our method is tested in three subcellular localization predictions for prokaryotic proteins and four subcellular localization predictions for eukaryotic proteins on Reinhardt's dataset. The results are then compared to several other methods. The total prediction accuracies of two tests are both 100% by a self-consistency test, and are 92.9% and 84.14% by the jackknife test, respectively. Esub8 also provides excellent results: the total prediction accuracies are 100% by a self-consistency test and 87% by the jackknife test.

Conclusions: Our method represents a different approach for predicting protein subcellular localization and achieved a satisfactory result; furthermore, we believe Esub8 will be a useful tool for predicting protein subcellular localizations in eukaryotic organisms.

Background

Over recent years the number of new genomes and protein sequences has increased dramatically. Therefore, reliable and efficient sequence analysis tools are urgently needed. The native subcellular localization of a protein is important for understanding gene/protein function. Aberrant subcellular localization of proteins has been observed in the cells of several diseases, such as cancer and Alzheimer's disease [1]. Therefore, knowing the protein's localization will be one important step identifying its function. Even if we already know a protein's function, information about protein localization may provide us insights into the specific enzyme pathway [2-5]. Experimental annotations of subcellular localization are often based on operational, biochemical definitions that can be error prone [1]. Therefore, predicting subcellular localization has become one of the central problems in bioinformatics.

Actually, some methods have been developed to quickly predict the subcellular localizations of proteins. Most of these methods can be classified into two classes: one is based on the N-terminal sorting signals [5] and the other is based on amino acid composition. One advantage of the former is a clear biological implication [6]. However, in large genome analysis projects, genes are usually automatically assigned and these assignments are often unreliable for the 5'-regions [7]. This can result in leader sequences being missed or only partially included, thereby causing problems for prediction algorithms depending on them. Therefore, most methods are based on the amino acid composition rather than the N-terminal sorting signals alone. Our method is also based on the amino acid composition.

Nakashima and Nishikawa [8] have indicated that intracellular and extracellular proteins differ significantly in their amino acid composition. There are already several algorithms based on amino acid compositions, such as least Mahalanobis distance [9-11], neural network [7], covariant discriminant algorithm [12,13], Markov Chain [14], and support vector machine [15,16]. Some researchers also consider combining other features together with amino acid composition. Feng and Zhang [17,18] proposed two methods: one combined the hydrophobic information, and the other combined Zp parameters. Recently, many novel methods have been developed based on new features. Gardy *et al.* developed a tool to predict protein subcellular localizations for Gram-negative bacteria, PSORT-B, which combined several methods together [19]. Rajesh and Burkhard developed a tool, LOC3D, to predict subcellular localizations for eukaryotic proteins of known three-dimensional (3D) structure [1]. Chou initially introduced the use of pseudo-amino-acid-composition to predict protein subcellular localization [20], and then Cai, Zhou and Chou developed several methods based on this new feature [13,21,22]. Functional domain composition was used by Chou and Cai [16,21] who also presented a new method incorporating gene ontology [22]. The results of above papers indicate that some of these new features can improve the prediction accuracy markedly, but a great shortcoming of these features is that it is difficult to obtain these features for new sequences, such as the functional domain composition and gene ontology.

We noted that in most of these methods using traditional amino acid composition to represent a protein, all the sequence-order information is neglected; consequently, the methods based on amino acid composition bear a bias of losing the sequence-order information. Chou firstly introduced a set of sequence-order-coupling numbers based on the physicochemical distance between amino acid to reflect the sequence order effect [23]. Actually this effect is a quasi-sequence-order effect. This paper takes into account sequence-order information by a different method. We also think that 1-v-1 multi-class SVM is better than 1-v-r SVM, so in this paper 1-v-1 SVM is used and the sequence-order information is also considered. We achieved excellent results: the total prediction accuracies of two tests on Reinhardt's dataset (predict three localizations for prokaryotic proteins and four localizations for eukaryotic proteins) are 100% by the self-consistency test, 92.9% and 84.14% by the jackknife test. Our method represents a different approach for predicting protein subcellular localization and achieved a satisfactory result. Our results show that the prediction accuracies are significantly improved. In this paper we also developed a tool, Esub8, to predict eight subcellular localizations for eukaryotic proteins: the total accuracies are 100% by the self-consistency test and 87% by the jackknife test. The results indicate Esub8 is a useful tool.

Results

Prediction accuracy

The prediction accuracies of subcellular localization for prokaryotic sequences on Reinhardt's dataset are shown in Table 1. The total accuracy by the self-consistency test reaches 100%. The total accuracy by the jackknife test reaches 92.9%. The prediction accuracies of subcellular localization for eukaryotic proteins Reinhardt's dataset are shown in Table 2. The total accuracy by the self-consistency test reaches 100%, the total accuracy by the jackknife test reaches 84.14%.

Table 1: Prediction accuracies of traditional subcellular localization for prokaryotic sequences with RBF kernel function

Location	Accuracy (%) (Self-consistency test)	Accuracy (%) (Jackknife test)
Extracellular	100	75.7 (75.7)
Periplasmic	100	81.2 (78.7)
Cytoplasmic	100	99 (97.5)
Total accruacy	100	92.9 (91.4)

Table 2: Prediction accuracies of traditional subcellular localization for eukaryotic sequences with RBF kernel function

Location	Accuracy (%) (Self-consistency test)	Accuracy (%) (Jackknife test)
Extracellular	100	86.5
Mitochondrial	100	67.6
Cytoplasmic	100	80
Nuclear	100	91.2
Total accruacy	100	84.14

Table 3: Prediction accuracies of Esub8 with RBF kernel function

Location	Accuracy (%) (Self-consistency test)	Accuracy (%) (Jackknife test)
Chloroplast	100	89.9
Cytoplasm	100	86.2
Extracellular	100	81.5
Golgi apparatus	100	68.2
Lysosome	100	85.0
Mitochondria	100	72.0
Nucleus	100	92.2
Peroxisome	100	72.6
Total accruacy	100	87

The prediction accuracies of Esub8 are shown in Table 3. The total accuracy by the self-consistency test reaches 100%, the total accuracy by the jackknife test reaches 87%. The kernel functions are all Radial Basis Functions (RBF). Esub8 and the other two traditional prediction programs, after cross-validation tests, all achieved optimal results with the same parameters: $C = 500$, $\gamma = 50$.

Comparison with other methods

In this section, our traditional localization results are compared with results obtained by other methods. These methods include Reinhardt and Hubbard's method using neural networks [7], Chou and Elrod's method using a covariant discriminant algorithm [12], Yuan's method based on the Markov Chain [14], Hua and Sun's method using a 1-v-r SVM method [15], Feng and Zhang's two methods using Bayesian discriminant function [17,18]. These methods are all based on Reinhardt and Hubbard's dataset [7], that is, all these methods used an identical dataset and their input vectors are all based on amino acid composition alone (Feng and Zhang's two methods are based on input vectors combining amino acid composition with other features). As shown in Table 4, for prokaryotic sequences, the total accuracy by the self-consistency test is about 10% higher than that of method 3 and about 2.3% higher than that of method 6. The total accuracy by the jackknife test is about 11.8% higher than that of method 2, 5.9% higher than that of method 3, 3.8% higher than that of method 4, 3.3% higher than that of method 7, 2.5% higher than that of method 6 and 1.5% higher than that of method 5. For eukaryotic sequences, other methods did not match the results of the self-consistency test; the total accuracy by the jackknife test is about 18.14% higher than that of method 2, 11.14% higher than that of method 4 and 4.74% higher than that of method 5. From these we know our method represents a different approach for predicting protein subcellular localization that achieved a satisfactory result.

Esub8 uses the same method to predict more rigorous localization (8 localizations) for eukaryotic proteins. From the data in Table 3, we know Esub8 is a satisfactory tool. The Institute of Bioinformatics, Tsinghua University also provided a web server, http://bioinfo.tsinghua.edu.cn/CoupleLoc/eu8.html, for eight localizations prediction of eukaryotic proteins, but the accuracies are unpublished.

Discussion

Subcellular localization of a new protein sequence is very important and fruitful for understanding its function, and predicting subcellular localization has become one of the central problems in bioinformatics. In this paper, we have developed a novel tool for protein eight subcellular localization predictions. We also test our method on Reinhardt's dataset. The proposed method differs from the

Table 4: Comparing the total accuracies with other 6 methods. From 1 to 7, the methods are our method, Reinhardt and Hubbard's method, Chou and Elrod's method, Yuan's method, Hua and Sun's method, Feng and Zhang's method 1 and method 2.

	1	2	3	4	5	6	7
P. S.	100	-	90	-	-	97.7	-
P. J.	92.9	81	87	89.1	91.4	90.4	89.6
E. S.	100	-	-	-	-	-	-
E. J.	84.14	66	-	73.0	79.4	-	-

P. S. denotes prokaryotic sequences prediction accuracies by self-consistency test, and then P. J., prokaryotic sequences prediction accuracies by jackknife test, E. S., eukaryotic sequences prediction accuracies by self-consistency test, E. J. eukaryotic sequences prediction accuracies by jackknife test. En dash denotes there is no result by the corresponding method.

existing method with the use of the 1-v-1 SVM and the order information of the protein sequence. The experimental results show that our method represents a different and satisfactory approach for predicting protein subcellular localization. Furthermore, our method has an advantage common to other methods based on amino acid composition: it is robust to errors in gene 5'-region annotation. We believe that Esub8 is a useful and efficient tool for protein localization prediction and is an important auxiliary tool for protein function prediction.

We have also found that the parameters of SVMs play an important role in the prediction results. RBF kernel function is better than linear kernel and polynomial kernel functions in solving this problem. After the cross-validation experiment, we obtain the optimal results with $C = 500$, $\gamma = 50$ both for Esub8 and for traditional subcellular localizations. We think SVMs with advanced kernel function will achieve better results. Combining our method with the method based on N-terminal sorting signal also will achieve better results. We also noted that looking for better features is very important. As described above, some new features were used, such as hydrophobic information [17], Zp parameters [18], pseudo-amino-acid-composition [20], and functional domain composition [16,21], and some of these methods achieved satisfactory results. More recently Keun-Joon and Minoru presented a method that used amino acid pairs as features based on SVM [25]. Another point that should be mentioned is that one can provide new datasets to be applied by this method.

Conclusions

In this paper, we proposed a novel tool to predict protein subcellular localizations for eukaryotic proteins based on amino acid composition alone. As a result, the total prediction accuracies of two traditional tests are both 100% by the self-consistency test, and are 92.9% and 84.14% by the jackknife test respectively. Esub8 also obtains excellent results: the total prediction accuracies are 100% by the self-consistency test and 87% by the jackknife test. As described above, our method represents a different approach for predicting protein subcellular localization and achieved a satisfactory result. We believe that Esub8 is a useful and efficient tool for protein localization prediction and is an important auxiliary tool for protein function prediction.

Methods

Materials

The training dataset used in Esub8 was downloaded at http://bioinfo.tsinghua.edu.cn/CoupleLoc, Institute of Bioinformatics, Tsinghua University. The dataset used to test our method on the three traditional subcellular localizations for prokaryotic proteins and four subcellular localizations for eukaryotic proteins is the same as that used by Reinhardt and Hubbard [7]. For more details, please contact the above authors. Table 5 shows the dataset used in Esub8, which includes 8305 eukaryotic sequences classified into 8 localization classes (chloroplast, cytoplasm, extracellular, golgi apparatus, lysosome, mitochondria, nucleus and peroxisome). Table 6 shows Reinhardt and Hubbard's dataset that includes 997 prokaryotic sequences, classified into three localization classes (extracellular, periplasmic and cytoplasmic), and 2427 eukaryotic sequences belonging to four localization classes (extracellular, mitochondrial, cytoplasmic and nuclear).

Table 5: The dataset used in Esub8.

Subcellular localization	Number of sequences
Chloroplast	1019
Cytoplasm	2088
Extracellular	595
Golgi apparatus	211
Lysosome	133
Mitochondria	644
Nucleus	3199
Peroxisome	116

Table 6: The final sequences in each location class of the dataset

Species	Subcellular localization	Number of sequences
Prokaryotic	Extracellular	107
	Periplasmic	202
	Cytoplasmic	688
Eukaryotic	Extracellular	325
	Mitochondrial	321
	Cytoplasmic	684
	Nuclear	1097

Feature vector

In many methods, the feature used to classify protein subcellular localizations is mainly amino acid composition [7-9,12,14,15,17,18]. In these papers, no matter how long the protein sequence is, the input vector is a twenty-dimensional vector because there are twenty kinds of amino acid in biological proteins. Each element in the feature vector denotes the presence frequency (or tendency) of an amino acid, so a feature vector can be represented by $\vec{x} \in R^{20}$. However, one drawback of this representation is that it neglects the order information of the protein sequence, that is, one cannot observe any amino acid order information from the feature vector. The order information may play an important role in protein subcellular localization.

In this paper, we present a novel approach for considering the sequence order information by dividing a protein sequence into two equal half sequences. For the first half sequence, we compute the amino acid composition to construct a 20D feature vector, and do the same with the second one. Then a forty-dimensional vector is constructed by combining the first 20D feature vector with the second one. Then the new feature vector can be represented by $\vec{x} \in R^{40}$. The results prove that our new 40D feature vector based on amino acid composition is better than 20D feature vector and then prove that amino acid order information plays an important role in protein subcellular localization.

Multi-Class SVM

SVM was introduced by Vapnik [26], and has been applied in many classification and regression problems. The standard SVM [26] was originally developed for dichotomic classification problems (binary classification). A classification problem usually involves training data and testing data that consist of some data instances. Each instance in training data contains one class label and one feature vector. The goal of SVM is to construct a classifier that classifies the data instances in the testing data.

For a binary classification problem, assume that we have a series of feature vectors x_i and class labels y_i (i = 1, 2... N, where N is the number of samples), where $x_i \in R^d$, $y_i \in \{+1, -1\}$. For protein sequences localization, the input vector dimension is 40, as described in the above section. The SVM requires the solution of the following optimization problem:

$$\text{Min } \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i$$

Subject to $y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i,\ \xi_i \geq 0.$ (1)

Here, feature vectors x_i are mapped into a higher dimensional space by the function $\phi(x) \in H$ and then SVM constructs an Optimal Separating Hyperplane (OSH), which maximizes the margin in the higher dimensional space. $C > 0$ is the penalty factor of the error term. Furthermore, $K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$ is called the kernel function. There are several typical kernel functions:

Polynomial kernel function: $K(x_i, x_j) = (x_i \bullet x_j + 1)^d$, (2)

Radia Basic Function (RBF): $K(x_i, x_j) = exp(-\gamma||x_i - x_j||^2)$, $\gamma > 0$, (3)

Sigmoid function: $K(x_i, x_j) = \tanh(\gamma\, x_i^T x_j + c)$ (4)

Here, d, γ and c are kernel parameters.

The multi-class classification problem is commonly solved by a decomposing and reconstructing procedure when the binary class SVM is implied. Protein subcellular localization is a multi-class problem, so we should decompose this problem into several binary classifications and then reconstruct them together. In this paper, we use the 1-v-1 SVM. For the 1-v-1 multi-class SVM, the decomposing method constructs all the possible binary machines from K-class training samples, each SVM being trained on only two out of all K classes. The usual reconstruction method is a parallel structure: when a new entry is presented, each binary learned machine provides one output concerning the classes involved in the training phase; then an algorithm interprets these two-class classifier outputs to determine the label to be assigned to the input. There exist several combinatorial algorithms for the outputs. Voting schemes are used in this paper because the output scale of a SVM is not robust. Since it depends on just the support vectors, voting schemes are more practical [27].

Implementation of the prediction system

In this paper, the 1-v-1 SVM was used to construct a protein subcellular localization system, Esub8, based on a 40D amino acid input vector. Esub8 is a program to classify one protein sequence into one of the eight classes. We also test our method on three traditional localizations for prokaryotic proteins and four localizations for eukaryotic proteins. Esub8 and the other programs were all written in Matlab using the software package, Osusvm, which was developed by Junshui Ma and Yi Zhao *et al*. based on SVMlight [28]. Our hardware platform is a PC running at 2.4 GHz. In traditional localizations, the self-consistency test can be finished in one minute; the jackknife test takes about two hours for all eukaryotic sequences and about 10 minutes for all prokaryotic sequences. In eight localizations, self-consistency can be finished in several minutes; the jackknife test takes about four days for all 8305 eukaryotic sequences. For Esub8, predicting the subcellular localization of an unknown sequence will take several seconds; hence, Esub8 is an efficient subcellular localization predication tool.

Self-consistency test and Jackknife test

Usually, the prediction results are evaluated by the self-consistency and jackknife tests. Although the sum-sampling test method is still widely used in biology literatures, the self-consistency and jackknife tests are more objective and rigorous, see Chou and Zhang's paper for a comprehensive discussion [29]. The former reflects the consistency of the prediction system, and the latter reflects the extrapolating effectiveness of the algorithm. When the self-consistency test is performed, the subcellular localizations of each protein in the dataset are in turn identified using the rule parameters derived from the training dataset. However, the prediction system parameters obtained by the self-consistency test are from the training dataset that includes the information of the later query protein. Since the same proteins are used to train the predictive system and test themselves, the error will be underestimated and the success rate will be enhanced, so a more reliable and rigorous test method, the jackknife test, is introduced. However, the self-consistency test is absolutely necessary because it reflects the self-consistency of the predictive system [30,31].

The jackknife test is the most effective and objective test method in statistical prediction. In the jackknife test, each protein in the dataset is singled out in turn as an independent test sample, and all the parameters of SVM are derived from training all the remaining proteins. In the process of jackknife tests, each protein has one chance to be the test sample, and for other tests this protein will be included in the training dataset.

Prediction system assessment

The total prediction accuracy is given by the following equations:

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

$$accuracy(i) = \frac{p(i)}{obs(i)}.$$

As described by Hua and Sun [15], N is the total number of sequences, k is the class number, $obs(i)$ is the number of sequences observed in localization i, and $p(i)$ is the number of correctly predicted sequences of localization i.

Authors' contributions

Qinghua Cui implemented the experiments in this study and wrote the draft of manuscript. Tianzi Jiang and Songde Ma supervised the whole process of this study, and Tianzi Jiang finalized the manuscript. Bing Liu participated in the validation of the study. All authors read and approved the final manuscript.

Acknowledgements

We are grateful to Dr. Elizabeth E. Budy for some language corrections. This work was partially supported by the Hundred Talents Programs of the Chinese Academy of Sciences, and the Natural Science Foundation of China, Grant No. 60172056 and 60121302.

References

1. Rajesh N, Burkhard R: **LOC3D: annotate sub-cellular localization for protein structures.** *Nucleic Acids Res* 2003, **13:**3337-3340.
2. Himmelreich R, Hilbert H, Plagens H, Pirkl E, Li BC, Herrmann R: **Complete sequence analysis of the genome of the bacterium Mycoplasma pneumoniae.** *Nucleic Acids Res* 1996, **24:**4420-4449.
3. Bult CJ, White O, Olsen GJ, Zhou L, Fleischmann RD, Sutton GG, Blake Ja, FitzGerald LM, Clayton RA, Gocayne JD, Kerlavage AR, Dougherty BA, Tomb JF, Adams MD, Reigh CI, Overbeek R, Kirkness EF, Weinstock KG, Merrick JM, Glodek A, Scott JL, Geoghagen NS, Venter JC: **Complete genome sequence of the methanogenic archaeon, Methanococcus jannaschii.** *Science* 1996, **273:**1058-1073.
4. NaKai K, Kanehisa M: **Expert system for predicting protein localization sites in Gram-negative bacteria.** *Proteins* 1991, **11:**95-110.
5. NaKai K, Kanehisa M: **A knowledge base for predicting protein localization sites in eukaryotic cells.** *Genomics* 1992, **14:**897-911.
6. Chou KC: **Prediction of protein signal sequences.** *Curr Protein Pep Sci* 2002, **3:**615-622.

7. Reinhardt A, Hubbard T: **Using neural networks for prediction of the subcellular localization of proteins.** *Nucleic Acids Res* 1998, **26:**2230-2236.
8. Nakashima H, Nishikawa K: **Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies.** *J Mol Biol* 1994, **238:**54-61.
9. Cedano J, Aloy P, Perez-Pons JA, Querol E: **Relation between amino acid composition and cellular localization of proteins.** *J Mol Biol* 1997, **266:**594-600.
10. Chou KC: **A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space.** *Proteins* 1995, **21:**319-344.
11. Chou KC, Zhang CT: **Predicting protein folding types by distance functions that make allowances for amino acid interactions.** *J Biol Chem* 1994, **269:**22014-22020.
12. Chou KC, Elrod D: **Protein subcellular localization prediction.** *Protein Eng* 1999, **12:**107-118.
13. Zhou GP, Doctor K: **Subcellular location prediction of apoptosis proteins.** *Proteins* 2003, **50:**44-48.
14. Yuan Z: **Prediction of protein subcellular localizations using Markov chain models.** *FEBS Lett* 1999, **451:**23-26.
15. Hua SJ, Sun ZR: **Support vector machine approach for protein subcellular localization prediction.** *Bioinformatics* 2001, **17:**721-728.
16. Chou KC, Cai YD: **Using functional domain composition and support vector machines for prediction of protein subcellular location.** *J Biol Chem* 2002, **277:**45765-45769.
17. Feng ZP, Zhang CT: **Prediction of the subcellular localization of prokaryotic proteins based on the hydrophobicity index of amino acids.** *Int J Biol Macromol* 2001, **28:**255-261.
18. Feng ZP, Zhang CT: **A graphic representation of protein sequence and predicting the subcellular localizations of prokaryotic proteins.** *Int J Biochem Cell Biol* 2002, **34:**298-307.
19. Gardy JL, Spencer C, Wang K, Ester M, Tusnady GE, Simon I, Hua S, deFays K, Lambert C, Nakai K, Brinkman FS: **PSORT-B: improving protein subcellular localization prediction for Gram-negative bacteria.** *Nucleic Acids Res* 2003, **13:**3613-3617.
20. Chou KC: **Prediction of protein cellular attributes using pseudo-amino-acid-composition.** *Proteins* 2001, **43:**246-255.
21. Cai YD, Chou KC: **Nearest neighbour algorithm for predicting protein subcellular by combining functional domain composition and pseudo-amino acid composition.** *Biochem Biophys Res Commun* 2003, **305:**407-411.
22. Chou KC, Cai YD: **Prediction and classification of protein subcellular location: sequence-order effect and pseudo amino acid composition.** *J Cell Biochem* 2003, **90:**1250-1260.
23. Pan YX, Zhang ZZ, Guo ZM, Feng GY, Huang ZD, He L: **Application of pseudo amino acid composition for predicting protein subcellular location: stochastic signal processing approach.** *J Protein Chem* 2003, **22:**395-402.
24. Chou KC, Cai YD: **A new hybrid approach to predict subcellular localization of proteins by incorporating Gene ontology.** *Biochem Biophys Res Commun* 2003, **311:**743-747.
25. Keun-Joon P, Minoru K: **Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs.** *Bioinformatics* 2003, **19:**1656-1663.
26. Vapnik V: **The Nature of Statistical Learning Theory.** *Springer: New York* 1998.
27. Angulo C, Parra X, Catala A: **K-SVCR. A support vector machine for multi-class classification.** *Neurocomputing* 2003, **55:**57-77.
28. Joachims T: **Making large-scale SVM learning practical.** In *Advances in Kernel Methods-Support Vector Learning* Edited by: Scholkopf B, Burges C, Smola A. MIT Press, Cambridge, MA; 1999:42-56.
29. Chou KC, Zhang CT: **Review: Prediction of protein structural classes.** *Crit Rev Biochem Mol Biol* 1995, **30:**275-349.
30. Cai YD: **Is it a paradox or misinterpretation.** *Proteins* 2001, **43:**336-338.
31. Zhou GP, Doctor K: **Subcellular location prediction of apoptosis proteins.** *Proteins* 2003, **50:**44-48.