Journal of Cellular Biochemistry 84:343–348 (2002)

Support Vector Machines for Prediction of Protein Subcellular Location by Incorporating Quasi-Sequence-Order Effect

Yu-Dong Cai,[1]* Xiao-Jun Liu,[2] Xue-biao Xu,[3] and Kuo-Chen Chou[4]

[1]Shanghai Research Centre of Biotechnology, Chinese Academy of Sciences, Shanghai, 200233, China
[2]Institute of Cell, Animal and Population Biology, University of Edinburgh, West Mains Road, Edinburgh EH9 3JT, United Kingdom
[3]Department of Computing Science, University of Wales, College of Cardiff, Queens Buildings, Newport Road, PO Box 916, Cardiff CF2 3XF, United Kingdom
[4]Computer-Aided Drug Discovery, Upjohn Laboratories, Kalamazoo, Michigan 49001-4940

Abstract Support Vector Machine (SVM), which is one class of learning machines, was applied to predict the subcellular location of proteins by incorporating the quasi-sequence-order effect (Chou [2000] Biochem. Biophys. Res. Commun. 278:477–483). In this study, the proteins are classified into the following 12 groups: (1) chloroplast, (2) cytoplasm, (3) cytoskeleton, (4) endoplasmic reticulum, (5) extracellular, (6) Golgi apparatus, (7) lysosome, (8) mitochondria, (9) nucleus, (10) peroxisome, (11) plasma membrane, and (12) vacuole, which account for most organelles and subcellular compartments in an animal or plant cell. Examinations for self-consistency and jackknife testing of the SVMs method were conducted for three sets consisting of 1,911, 2,044, and 2,191 proteins. The correct rates for self-consistency and the jackknife test values achieved with these protein sets were 94 and 83% for 1,911 proteins, 92 and 78% for 2,044 proteins, and 89 and 75% for 2,191 proteins, respectively. Furthermore, tests for correct prediction rates were undertaken with three independent testing datasets containing 2,148 proteins, 2,417 proteins, and 2,494 proteins producing values of 84, 77, and 74%, respectively. J. Cell. Biochem. 84: 343–348, 2002.

Key words: Support Vector Machines; protein subcellular location; quasi-sequence-order-effect

Knowledge of the subcellular location of a given protein is very important for understanding its function [Chou and Elrod, 1999a,b; Chou, 2000a,b]. However, due to the rapid increase in the number of protein sequences, it is time consuming and costly to determine their subcellular location solely by laboratory experiments. Accordingly, it is highly desirable to develop an effective algorithm that can quickly and accurately predict protein subcellular location. Many efforts have been made in this regard [Nakashima and Nishikawa, 1994; Cedano et al., 1997; Reinhardt and Hubbard, 1998; Chou and Elrod, 1998, 1999b].

The covariant discriminate algorithm [Chou and Elrod, 1999b], which was developed from the least Mahalanobis distance algorithm [Chou, 1995], has proved particularly successful. However, all these methods were based on the amino acid composition and ignored sequence order. Such prediction approaches might be approximately rational, but the success rates would be limited. To improve prediction quality, Chou [2000b] proposed a new method in which the covariant discriminate algorithm [Chou and Elrod, 1999b] was augmented to incorporate the quasi-sequence-order effect. The new method allows using both the sequence-order-coupling numbers that reflect the sequence order effect and amino acid composition in order to improve prediction quality. The incorporation of the quasi-sequence-order effect for prediction of protein subcellular location is one step forward in this area. In the current paper, by incorporating sequence order effect, we have employed Vapnik's Support Vector Machine (SVM) [Vapnik, 1995] to predict protein subcellular location.

*Correspondence to: Yu-Dong Cai, Biomolecular Sciences Department, UMIST, P.O. Box 88, Manchester, M60 1QD, UK. E-mail: y.cai@umist.ac.uk

Received 23 April 2001; Accepted 5 September 2001

DOI 10.1002/jcb.10030

SUPPORT VECTOR MACHINE

Support Vector Machine is a class of learning machines based on statistical learning theory. The basic idea of applying SVM to pattern classification can be stated briefly: First, map the input vectors into one feature space (possible with a higher dimension), either linearly or non-linearly, which is relevant to the selection of the kernel function. Then, within the feature space from the first step, seek an optimized linear division, i.e., construct a hyperplane which separates two classes (this can be extended to multi-class). SVM training always seeks a global optimized solution and avoids over-fitting, so it has the ability to deal with a large number of features. A complete description to the theory of SVMs for pattern recognition is in Vapnik's book [Vapnik, 1998]. SVMs have been used in a range of problems including drug design [Burbidge et al., 2000], image recognition, and text classification [Joachims, 1998].

In this paper, we apply Vapnik's SVM [Vapnik, 1995] for predicting protein subcellular location. We download the SVMlight, which is an implementation (in C Language) of SVM for the problem of pattern recognition. The optimization algorithm used in SVM light has been described [Joachims, 1999a,b]. The code has been used in text classification and image recognition [Joachims, 1998].

Suppose we are given a set of samples, i.e., a series of input vectors

$$X_i \in R^d \quad (i = 1, \ldots, N)$$

with corresponding labels $y_i \in \{+1, -1\}\,(i = 1, \ldots, N)$ where -1 and $+1$ are used to stand respectively for the two classes. The goal here is to construct one binary classifier or derive one decision function which has small probability of misclassifying a future sample (from the available samples). Both the basic linear separable case and the most useful linear non-separable case for most real life problems are considered here.

Linear Separable Case

In this case, there exists a separating hyper plane whose function is $\vec{W} \bullet \vec{X} + b = 0$, which implies:

$$y_i(\vec{W} \bullet \vec{x}_i + b) \geq 1, \quad i = 1, \ldots, N$$

By minimizing $\frac{1}{2}\|\vec{W}\|^2$ subject to this constraint, the SVM approach tries to find a unique separating hyperplane. Here $\|\vec{W}\|^2$ is the Euclidean norm of $\vec{w}$, which maximizes the distance between the hyper plane (Optimal Separating Hyperplane or OSH in Cortes and Vapnik, [1995]) and the nearest data points of each class. The classifier is called the largest margin classifier.

By introducing Lagrange multipliers α_i, using the Karush-Kuhn-Tucker (KKT) conditions and the Wolfe dual theorem of optimization theory, the SVM training procedure amounts to solving the following convex QP problem:

$$Max : \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j \cdot y_i y_j \cdot \vec{X}_i \bullet \vec{X}_j$$

subject to the following two conditions:

$$\alpha_i \geq 0$$

$$\sum_{i=1}^{N} \alpha_i y_i = 0, \quad i = 1, \ldots, N$$

The solution is a unique globally optimized result having the following expansion:

$$\vec{W} = \sum_{i=1}^{N} y_i \alpha_i \cdot \vec{x}_i$$

Only if the corresponding $\alpha_i > 0$, these $\vec{x}_i$ are called Support Vectors.

When a SVM is trained, the decision function can be written as:

$$f(\vec{x}) = \mathrm{sgn}\left(\sum_{i=1}^{N} y_i \alpha_i \cdot \vec{x} \bullet \vec{x}_i + b\right)$$

sgn( ) appears in the above formula as the given sign function.

Linear non-separable case. Two important techniques needed for this case are given respectively as below:

(i) "soft margin" technique.

In order to allow for training errors, Cortes and Vapnik [1995] introduced slack variables:

$$\xi_i > 0, \quad i = 1, \ldots, N$$

and relaxed separation constraint is given as:

$$y_i(\vec{w} \bullet \vec{x}_i + b) \geq 1 - \xi_i, \quad (i = 1, \ldots, N)$$

and the OSH can be found by minimizing

$$\frac{1}{2}\|\vec{w}\|^2 + C\sum_{i=1}^{N} \xi_i$$

instead of $\frac{1}{2}\|\vec{w}\|^2$ for the above two constraints in "Linear separation case," where c is a regularization parameter used to decide a trade-off between the training error and the margin.

(ii) "kernel substitution" technique

SVM performs a nonlinear mapping of the input vector $\vec{x}$ from the input space R^d into a higher dimensional Hilbert space, where the mapping is determined by the kernel function. Then like in "Linear separation case," it finds the OSH in the space H corresponding to a nonlinear boundary in the input space.

Two typical kernel functions are listed below:

$$K(\vec{x_i},\vec{x_j}) = (\vec{x_i} \bullet \vec{x_j} + 1)^d$$

$$K(\vec{x_i},\vec{x_j}) = \exp(-r\|\vec{x_i} - \vec{x_j}\|^2)$$

where the first one is called the *polynomial kernel function of degree d*, which will eventually revert to the linear function when $d = 1$, the latter one, is called the RBF (Radial Basic Function) kernel. Finally, for the selected kernel function, the learning task amounts to solving the following QP problem,

$$Max : \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i\alpha_j \cdot y_i y_j \cdot K(\vec{X_i} \bullet \vec{X_j})$$

subject to:

$$0 \le a_i \le C$$

$$\sum_{i=1}^{N} \alpha_i y_i = 0, \quad i = 1,\ldots,N$$

and the form of the decision function is

$$f(\vec{x}) = \mathrm{sgn}\left(\sum_{i=1}^{N} y_i\alpha_i \cdot K(\vec{x},\vec{x_i}) + b\right)$$

For a given data set, only the kernel function and the regularity parameter C must be selected to specify one SVM.

Training and Prediction of Protein Subcellular Location

Following the procedures and rationale as given by Chou and Elrod [1999b], the proteins are classified into the following 12 groups: (1) chloroplast, (2) cytoplasm, (3) cytoskeleton, (4) endoplasmic reticulum, (5) extracell, (6) Golgi apparatus, (7) lysosome, (8) mitochondria, (9) nucleus, (10) peroxisome, (11) plasma membrane, and (12) vacuole, which have covered almost all the organelles and subcellular compartments in an animal or plant cell.

Following the procedures as given by Chou [2000], the sequence order effect can be approximately reflected through a set of sequence-order-coupling numbers as defined below:

Suppose a protein chain of L amino acid residues:

$$R_1R_2R_3R_4R_5R_6R_7\cdots R_L,$$

then the sequence order effect can be approximately reflected through a set of sequence-order-coupling numbers as defined below:

$$\begin{cases} \tau_1 = \dfrac{1}{L-1}\displaystyle\sum_{i=1}^{L-1} J_{i,i+1} \\ \tau_2 = \dfrac{1}{L-2}\displaystyle\sum_{i=1}^{L-2} J_{i,i+2} \\ \tau_3 = \dfrac{1}{L-3}\displaystyle\sum_{i=1}^{L-3} J_{i,i+3} \\ \cdots \\ \tau_\varphi = \dfrac{1}{L-\varphi}\displaystyle\sum_{i=1}^{L-\varphi} J_{i,i+\varphi} \end{cases}, (\varphi < L) \qquad (1)$$

where τ_1 is called the 1st-rank sequence-order-coupling number that reflects the coupling mode between all the most contiguous residues along a protein sequence, τ_2 is the 2nd-rank sequence-order-coupling number that reflects the coupling mode between all the 2nd most contiguous residues, and so forth. In Eq. [1], the coupling factor $J_{i,j}$ is a function of amino acids R_i and R_j, we choose

$$J_{i,j} = D^2(R_i, R_j), \qquad (2)$$

where $D(R_i, R_j)$ is the physicochemical distance from amino acid R_i to amino acid R_j that was derived based on the residue properties of hydrophobicity, hydrophilicity, polarity, and side chain volume, see Joachims [1999b].

Suppose there are N proteins forming a set S, which is the union of m subsets; i.e.,

$$S = S_1 \cup S_2 \cup S_3 \cup S_4 \cup \cdots \cup S_m \qquad (3)$$

Each subset is composed of proteins with a same subcellular location. Its size is given by $n_\xi(\xi = 1,\ 2, 3, \ldots, m)$, where n_ξ represents the number of proteins in the subset S_ξ.

The kth protein in the subset S_ξ should now be described by

$$X_k^\xi = \begin{bmatrix} x_{k,1}^\xi \\ x_{k,2}^\xi \\ \vdots \\ x_{k,20+\varphi}^\xi \end{bmatrix}, (k = 1, 2, \ldots, n_\xi; \xi = 1, 2, \ldots, m), \tag{4}$$

where

$$x_{k,u}^\xi = \begin{cases} \dfrac{f_{k,u}^\xi}{\sum_{j=1}^{20} f_{kj}^\xi + w \sum_{q=1}^{\varphi} \tau_{k,q}^\xi}, (1 \le u \le 20) \\ \dfrac{w\tau_{k,u-20}^\xi}{\sum_{j=1}^{20} f_{kj}^\xi + w \sum_{q=1}^{\varphi} \tau_{k,q}^\xi}, (20+1 \le u \le 20+\varphi) \end{cases} \tag{5}$$

where $f_{k,j}^\xi$ is the normalized occurrence frequency of the 20 amino acids in the kth protein in subset S_ξ, $\tau_{k,q}^\xi$ is the qth-rank sequence-order-coupling number computed according to Eqs. [1] and [2] for the kth protein in subset S_ξ, and w is the weight factor for the sequence-order effect. Here, we choose $w = 0.1$. As we can see from Eqs.[4] and [5], the first 20 components reflect the effect of the amino acid composition, while the components from $20 + 1$ to $20 + \varphi$ reflect the effect of sequence order.

In this research, $\varphi = 13$, therefore, a protein can be represented by a point or a vector in a 33-D space. These are taken as the input of the SVM.

The computations were carried out on a Silicon Graphics IRIS Indigo workstation (Elan 4000).

In this research, for the SVM, the width of the Gaussian RBFs is selected as that which minimized an estimate of the VC-dimension. The parameter C that controls the error-margin tradeoff is set at 100. After being trained, the hyperplane output by the SVM was obtained. This indicates that the trained model, i.e., hyperplane output which is including the important information, has the function of identifying the subcellular location.

In this research, first the self-consistency and jackknife tests (leave-one-out) of the method were performed, and later the calculation was extended to deal with an independent data set. As a result, high rates of correct prediction were obtained in all three tests.

RESULTS AND DISCUSSION

Success Rates of Self-Consistency and Jackknife Test of SVMs

In this study, the examination for the self-consistency of the SVMs method was tested for the three sets from Table I [Chou, 2000b]: 1,911 proteins (chloroplast: 145, cytoplasm: 571, extracell: 224, nucleus: 272, plasma membrane: 699); 2,044 proteins (chloroplast: 145, cytoplasm: 571, endoplasmic reticulum: 49, extracell: 224, mitochondria: 84, nucleus: 272, plasma membrane: 699); 2,191 proteins (chloroplast: 145, cytoplasm: 571, cytoskeleton: 34, endoplasmic reticulum: 49, extracell: 224, Golgi apparatus: 25, lysosome: 37, mitochondria: 84, nucleus: 272, peroxisome: 27, plasma membrane: 699, vacuole: 24). As a result, the correct rate of self-consistency reached 94, 92, and 89% for 1,911, 2,044, and 2,191 proteins, respectively, which showed that after being trained, the SVMs

TABLE I. Self-Consistency and Jackknife Test Results for the 2,191 Proteins

	Rate of correct prediction for each subcellular location			
	(1)Chloroplast	(2)Cytoplasm	(3)Cytoskeleton	(4)Endoplasmic reticulum
Self-consistency	121/145 = 84%	557/571 = 98%	19/34 = 56%	31/49 = 63%
Jackknife test	82/145 = 57%	504/571 = 88%	15/34 = 44%	15/49 = 31%
	(5)Extracell	(6)Golgi apparatus	(7)Lysosome	(8)Mitochondria
Self-consistency	180/224 = 80%	10/25 = 40%	32/37 = 86%	59/84 = 70%
Jackknife test	127/224 = 57%	3/25 = 12%	20/37 = 54%	35/84 = 42%
	(9)Nucleus	(10)Peroxisome	(11)Plasma membrane	(12)Vacuole
Self-consistency	243/272 = 89%	4/27 = 15%	675/699 = 97%	16/24 = 67%
Jackknife test	198/272 = 73%	1/27 = 4%	636/699 = 91%	6/24 = 25%

method had grasped the complicated relationship between the amino acid composition, sequence order, and subcellular location.

The above results indicate a very good self-consistency of the current prediction method. Below we describe the cross-validation tests. The single independent data set test, sub-sampling test, and jackknife test are the three methods ordinarily used for cross-validation. Of the three cross validation methods, the jackknife test is deemed most effective and objective [Chou, 2000b; Cai, 2001]. During the jackknifing process, both the training and testing datasets are actually open, and a protein will in turn move from one to the other [Chou, 2000b]. As a result, the correct rate reaches 83, 78, and 75% for 1,911, 2,044, and 2,191 protein sets, respectively.

The results of both self-consistency and jackknife tests for 12 subcellular locations of the 2191 protein set is provided in Table I. The success rates by jackknife test are generally lower than those by the self-consistency test. In the self-consistency test, each protein sequence from a data set is predicted using the rule parameters derived from the same data set, the so-called training data set. As a consequence, the parameters derived from the training data set include the information of a protein later plugged back in the test. This will certainly give a somewhat optimistic error estimate because the same proteins are used to derive the prediction rules and to test themselves. Nevertheless, such a re-substitution test is absolutely necessary because it reflects the self-consistency of a prediction method, especially for its algorithm part. A prediction algorithm certainly cannot be deemed as a good one if its self-consistency is poor. In other words, the re-substitution test is necessary but not sufficient for evaluating a prediction method. As a complement, a cross-validation test for an independent data set is needed because it can reflect the extrapolating effectiveness of a prediction method. This is important especially for checking the validity of a training data set: whether it contains sufficient information to reflect all the important features concerned so as to yield a high success rate in practical application. However, how to perform the cross-validation is a subtle problem. It is well known that the single independent data set test, sub-sampling test, and jackknife test are the three methods often used for cross-validation (see a review article by Chou and Zhang [1995] for a comprehensive discussion about this). Of the three cross validation methods, the jackknife test is deemed as the most effective and objective one [Cai, 2001; Zhou and Assa-Munt, 2001]. During jackknifing, each protein in a data set is in turn singled out as a tested protein and all the rule-parameters are computed using the remaining proteins without including this one. In other words, the protein cellular location of each protein is predicted by the rules derived using all other proteins except the one that is being predicted. In the process of jackknife test, both the training data set and testing data set are actually open, and a protein will in turn move from each to other [Chou et al., 1998]. The cross-validation through such a jackknife approach is much more objective and rigorous than the other two test approaches. In view of this, it is also clear why the success rate by jackknife test for small subset, such as Golgi apparatus and lysosome, is reduced more remarkably than those of a larger subset. Therefore, the information loss resulting from jackknifing will have greater impact on the small subsets than the larger ones. It is anticipated that the jackknife rates for the small subsets can be improved by adding into them more new proteins that have been found belonging to the locations defined by these subsets.

Success Rate of Correct Prediction of the SVMs for an Independent Dataset

The corresponding three independent testing datasets contain 2,148 proteins (chloroplast: 112, cytoplasm: 761, extracell: 95, nucleus: 418, plasma membrane: 762); 2,417 proteins (chloroplast: 112, cytoplasm: 761, endoplasmic reticulum: 106, extracell: 95, mitochondria: 163, nucleus: 418, plasma membrane: 762); 2,494 proteins (chloroplast: 112, cytoplasm: 761, cytoskeleton: 19, endoplasmic reticulum: 106, extracell: 95, Golgi apparatus: 4, lysosome: 31, mitochondria: 163, nucleus: 418, peroxisome: 23, plasma membrane: 762, vacuole: 0), respectively. The correct prediction rates reach 84, 77, and 74% for the 2,148, 2,417, and 2,494 protein sets, respectively.

The datasets used here were generated by strictly following certain screening procedures to minimize the possibility of any two similar sequences occurring in a same category. In addition, the sequence matches performed between all members in each category of proteins

thus obtained have indicated that most pairs have very low sequence identity ($< 20\%$). The average sequence identity in each category is smaller than 12%. The number of pairs having high sequence identity ($> 90\%$) is very small. The percentages of pairs having $> 90\%$ sequence identity in the chloroplast, cytoplasm, cytoskeleton, endoplasmic reticulum, extracellular, Golgi apparatus, lysosome, mitochondria, nucleus, peroxisome, plasma membrance, and vacuole subsets are 0.12, 0.04, 0.036, 0.034, 0.02, 0, 0, 0, 0.01, 0.057, 0.01, and 1.1%, respectively. Obviously, such a small fraction of high-sequence identity proteins cannot be the origin of the high rates obtained by SVM.

The SWISS-PROT codes for all the proteins studied here are from Appendix A of Chou and Elrod [1999b]. From these protein codes, all the protein sequences employed in this study can be retrieved from the SWISS-PROT data bank.

CONCLUSION

The above results, together with those obtained by the covariant discriminant prediction algorithm [Chou and Elrod, 1998, 1999b; Chou, 2000b], indicate that the cellular location of a protein can be predicted with reasonable accuracy. It is anticipated that the covariant discriminant algorithm [Chou and Elrod, 1998, 1999b; Chou, 2000b] and the SVMs, if complemented with each other, will become a powerful tool for predicting the subcellular locations of proteins, and hence facilitate the systematic analysis of genome data.

REFERENCES

Burbidge R, Trotter M, Holden S, Buxton B. 2000. Drug design by machine learning: support vector machine for pharmaceutical data analysis. Proceedings of the AISB'00 Symposium on Artificial Intelligence in Bioinformatics. 1–4.

Cai YD. 2001. Is it a paradox or misinterpretation? Proteins: Funct Genet 43(3):336–338.

Cedano J, Aloy P, Perez-pons JA, Querol E. 1997. Relation between amino acid composition and cellular location of proteins. J Mol Biol 266(3):594–600.

Chou KC. 1995. A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. Proteins 21(4):319–344.

Chou KC. 2000a. Review: prediction of protein structural classes and subcellular locations. Curr Protein Peptide Sci 1:171–208.

Chou KC. 2000b. Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. Biochem Biophys Res Commun 278(2):477–483.

Chou KC, Elrod DW. 1998. Using discriminant function for prediction of subcellular location of prokaryotic proteins. Biochem Biophys Res Commun 252(1):63–68.

Chou KC, Elrod DW. 1999a. Prediction of membrane protein types and subcellular locations. Proteins: Structure Funct Genet 34(1):137–153.

Chou KC, Elrod DW. 1999b. Protein subcellular location prediction. Protein Eng 12(2):107–118.

Chou KC, Zhang CT. 1995. Prediction of protein structural classes. Biochem Mol Biol [Critical Review] 30(4):275–349.

Chou KC, Liu W, Maggiora GM, Zhang CT. 1998. Prediction and classification of domain structural classes. Proteins: Structure Funct Genet 31:97–103.

Cortes C, Vapnik V. 1995. Support vector networks. Machine Learning 20:273–293.

Joachims T. 1998. Text categorization with support vector machines: learning with many relevant features. Proceedings of the European Conference on Machine Learning, Springer.

Joachims T. 1999a. 11 in: Making large-Scale SVM learning practical. advances in kernel methods-support vector learning, Schölkopf B, Burges C, Smola A, editors. MIT Press.

Joachims T. 1999b. Transductive inference for text classification using support vector machines. International Conference on Machine Learning (ICML).

Nakashima H, Nishikawa K. 1994. Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. J Mol Biol 238(1):54–61.

Reinhardt A, Hubbard T. 1998. Using neural networks for prediction of the subcellular location of proteins. Nucleic Acids Res 26(9):2230–2236.

Vapnik V. 1995. The nature of statistical learning theory. Springer.

Vapnik V. 1998. Statistical learning theory. New York: Wiley-Interscience.

Zhou G, Assa-Munt N. 2001. Some insights into protein structural class prediction. Proteins: Structure Funct Genet 44:57–59.