Internet Electronic Journal of Molecular Design **2002**, *1*, 219–226 ISSN 1538–6414
BioChem Press http://www.biochempress.com

Inter*net* Electronic Journal of
Molecular Design

April 2002, Volume 1, Number 4, Pages 219–226

Editor: Ovidiu Ivanciuc

Support Vector Machines for Predicting Membrane Protein Types 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, U.K.
[3] Department of Computing Science, University of Wales, College of Cardiff, Queens Buildings, Newport Road, PO Box 916, Cardiff CF2 3XF, U.K.
[4] Computer–aided drug discovery, Upjohn Laboratories, Kalamazoo, Michigan 49007–4940, USA

Received: January 15, 2002; Revised: February 14, 2002; Accepted: February 27, 2002; Published: April 30, 2002

Citation of the article:
Y.–D. Cai, X.–J. Liu, X. Xu, and K.–C. Chou, Support Vector Machines for Predicting Membrane Protein Types by Incorporating Quasi–Sequence–Order Effect, *Internet Electron. J. Mol. Des.* **2002**, *1*, 219–226, http://www.biochempress.com.

Inter*net* Electronic Journal of Molecular Design

BioChem Press
http://www.biochempress.com

Support Vector Machines for Predicting Membrane Protein Types 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, U.K.

[3] Department of Computing Science, University of Wales, College of Cardiff, Queens Buildings, Newport Road, PO Box 916, Cardiff CF2 3XF, U.K.

[4] Computer–aided drug discovery, Upjohn Laboratories, Kalamazoo, Michigan 49007–4940, USA

Received: January 15, 2002; Revised: February 14, 2002; Accepted: February 27, 2002; Published: April 30, 2002

Internet Electron. J. Mol. Des. **2002**, *1* (4), 219–226

Abstract

Membrane proteins can be classified among the following five types: (1) type I membrane protein; (2) type II membrane protein; (3) multipass transmembrane proteins; (4) lipid chain–anchored membrane proteins; (5) GPI–anchored membrane proteins. A new learning machine, the Support Vector Machine, is applied for predicting the type of a given membrane protein by incorporating the quasi–sequence–order effect. High success rates were obtained by the self–consistency test (2030/2059 = 99%), jackknife test (1696/2059 = 82%), and independent data test (2305/2625 = 88%).

Keywords. Support Vector Machines; membrane protein types; quasi–sequence–order effect.

1 INTRODUCTION

A cell is enclosed by the plasma membrane (cell envelope). Inside the cell there are various organelles such as the endoplasmic reticulum, Golgi apparatus, mitochondria, and other membrane–bound organelles. Although the basic structure of biological membranes is provided by the lipid bilayer, most of the specific functions are carried out by the membrane proteins. Among membrane proteins, some of them are transmembrane proteins. They contain one or more transmembrane segments with one or more hydrophobic segments to ensure stable association with the hydrophobic interior of the membrane, and hence is relatively easily discriminated from non–membrane proteins [1]. The other membrane proteins are anchored membrane proteins. They do not have the

* Correspondence author; present address: Biomolecular Sciences Department, UMIST, P.O. Box 88, Manchester, M60 1QD, U.K.; E–mail: y.cai@umist.ac.uk.

hydrophobic membrane spanning portions, but they have a consensus sequence motif at either the N– or C– terminus. So they also can be relatively easily discriminated from non–membrane proteins [2,3]. In this paper, the discrimination is confined within the scope of membrane proteins only. This is because membrane proteins can be reliably distinguished by using existing methods, as elaborated by many previous investigators [4,5].

The way in which a membrane–bound protein is associated with the lipid bilayer usually reflects the function of the protein. The transmembrane proteins, for example, can function on both sides of the membrane and transport molecules from one side to other side, whereas the proteins that are associated with one side of the lipid monolayer or a protein domain can only function on that side. It is clear that the function of new proteins can be determined if an effective algorithm is available to predict their types. Chou and Elrod [4] classified membrane proteins into five different types. These authors proposed a covariant discriminant algorithm [4] to predict the types of membrane proteins. Recently, Cai *et al*. [6] applied neural network to this problem.

To improve the prediction quality, Chou proposed [7] a new method in which the covariant discriminate algorithm was augmented to incorporate the quasi–sequence–order effect. The new method uses the amino acid composition and the sequence–order–coupling numbers (reflecting the sequence order effect) in order to improve the prediction quality. The incorporation of the quasi–sequence–order effect for the prediction of the types of membrane proteins is one step forward in this area. Encouraged by the positive impact of including the quasi–sequence–order effect, we try to apply Vapnik's Support Vector Machine [8,9] to approach this problem.

2 SUPPORT VECTOR MACHINE

Support Vector Machine (SVM) is one type of learning machines based on statistical learning theory. The basic idea of applying SVM to pattern classification can be stated briefly as follows. First, map the input vectors into one feature space (possible with a higher dimension), either linearly or non–linearly, which is relevant with 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 problems). 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 [9]. SVMs have been used in a range of bioinformatics problems including protein fold recognition [10,11]; protein–protein interactions prediction [12]; prediction of protein subcellular location [13–15]; protein secondary structure prediction [16].

In this paper, we apply Vapnik's Support Vector Machine [8,9] for predicting the types of membrane proteins. We have used the SVMlight, which is an implementation (in the C language) of

SVM for the problem of pattern recognition. The optimization algorithm used in SVMlight can be found in [17].

Suppose we are given a set of samples, *i.e.* a series of input vectors $X_i \in R^d (i = 1,..., N)$, with corresponding labels $y_i \in \{+1,-1\} (i = 1,..., N)$, 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 from the available samples, which has small probability of misclassifying a future sample. Both the basic linear separable case and the most useful linear non–separable case (for most real life problems) are considered here.

2.1 The Linear Separable Case

In this case, there exists a separating hyperplane whose function is $\vec{W} \cdot \vec{X} + b = 0$, which implies the following:

$$y_i (\vec{W} \cdot \vec{x}_i + b) \geq 1, i = 1,..., N$$

By minimizing $0.5 \| \vec{W} \|^2$ 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 hyperplane (Optimal Separating Hyperplane or OSH in [18]) 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:

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

subject to the following two conditions:

$$\alpha_i \geq 0$$

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

The solution is a unique globally optimized result can be shown 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:

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

where $\text{sgn}(\ )$ in the above formula is the sign function.

2.2 The Linear Non–Separable Case

Two important techniques needed for this case are given respectively as below.

Soft margin technique. In order to allow for training errors, ref [18] introduced slack variables:

$$\xi_i > 0, i = 1,..., N$$

The relaxed separation constraint is given as:

$$y_i(\vec{W} \cdot \vec{X}_i + b) \geq 1 \ , \ (i = 1,..., N)$$

and the OSH can be found by minimizing:

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

instead of $0.5\|\vec{W}\|^2$ for the above two constraints. In the above equation C is a regularization parameter used to decide a trade–off between the training error and the margin.

Kernel substitution technique. SVM performs a nonlinear mapping of the input vector $\bar{x}$ from the input space R^d into a higher dimensional Hilbert space, where the mapping is determined by the kernel function. Then like in the linear separable case, it finds the OSH in the space H corresponding to a non–linear boundary in the input space. Two typical kernel functions are listed below:

$$K(\vec{x}_i, \vec{x}_j) = (\vec{x}_i \cdot \vec{x}_j + 1)^d$$

$$K(\vec{x}_i, \vec{x}_j) = \exp(-r \| \vec{x}_i - \vec{x}_j \|^2)$$

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

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

subject to:

$$0 \leq a_i \leq C$$

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

where the form of the decision function is

$$f(\vec{x}) = \text{sgn}(\sum_{i=1}^{N} y_i \alpha_i \cdot K(\vec{x}, \vec{x}_i) + b)$$

For a given data set, only the kernel function and the regularity parameter *C* must be selected.

3 TRAINING AND PREDICTION OF MEMBRANE PROTEIN TYPES

Following the procedures and rationale as given in [4], membrane proteins are classified into the following 5 types: (1) type I membrane protein; (2) type II membrane protein; (3) multipass transmembrane proteins; (4) lipid chain–anchored membrane proteins; and (5) GPI–anchored membrane proteins. Following the Chou's sequence–order–coupling procedure [7], the sequence order effect of a protein chain can be approximately reflected through a set of sequence–order–coupling numbers as defined below. Suppose a protein chain of L amino acid residues: $R_1R_2R_3R_4R_5R_6R_7\cdots R_L$. The sequence order effect can be approximately reflected through a set of sequence–order–coupling numbers as defined below:

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

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, given by:

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

where $D(R_i, R_j)$ is the physico–chemical 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 [7]. Suppose there are N proteins forming a set S, which is the union of m subsets:

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

Each subset is composed of proteins with a same type. 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 k th protein in the subset S_ξ may be described by

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

with

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

where $f_{k,j}^{\xi}$ is the normalized occurrence frequency of the 20 amino acids in the k^{th} protein in subset S_{ξ}, $\tau_{k,q}^{\xi}$ is the q^{th}–rank sequence–order–coupling number computed according to Eqs. [1] and [2] for the k^{th} 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 reflects 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, φ = 20, therefore, a protein can be represented by a point or a vector in a 40–D space. These are taken as the input of the SVM. As an example, 41BB_HUMAN can be computed and represented by a vector in 40–D space: (4.31 9.80 4.31 5.49 6.27 8.24 0.78 2.75 5.10 8.24 1.18 4.71 7.06 4.71 6.27 8.63 6.67 4.31 0.39 0.78 7.61 8.35 8.34 8.43 9.38 8.98 8.98 8.38 8.95 8.24 8.57 8.64 9.24 9.22 8.99 8.60 9.59 8.81 8.13 8.37). The computations were carried out on a Silicon Graphics IRIS Indigo workstation (Elan 4000). Also for the SVM, the width of the Gaussian RBFs (in this paper, we use the default value in SVMlight) is selected as that which minimized an estimate of the VC–dimension. The parameter *C* that controls the error–margin tradeoff is set at 1000. After being trained, the hyperplane output by the SVM was obtained. The SVM method applies to two–class problems. In this paper, for the five–class problems, we have used a simple and effective method: "one–against–others" method [10] to transfer it into two–class problems. We first test the self–consistency (model calibration) and leave–one–out cross–validation (jackknife test) of the method, followed by testing the method by prediction of an independent dataset. As a result, the rates of self–consistency, cross–validation and prediction were quite high.

4 RESULTS AND DISCUSSION

In this research, the examination for the self–consistency of the SVM (support vector machines) method was tested for the dataset from Chou and Elrod [4] that consists of 435 type I membrane protein, 152 type II membrane protein, 1311 multipass transmembrane proteins, 51 lipid chain–anchored membrane proteins, and 110 GPI–anchored membrane proteins). As a result, the success rates reach 98%, 94%, 99%, 100% and 95% for type I membrane protein, type II membrane protein, multipass transmembrane proteins, lipid chain–anchored membrane proteins and GPI–anchored membrane proteins, respectively. The overall success rate reaches 99%, which shows that after being trained, the SVM model has grasped the complicated relationship between the amino acid

composition and the types of membrane proteins.

Next, we examined the prediction quality by the jackknife test. During the process of the jackknife test, the training and testing datasets are actually open, and a protein will in turn move from each to the other. As a result, the overall success rate reaches 82%.

The combination of the self–consistency test and jackknife test as conducted above is thought the most effective method for examining the power of a method of statistical prediction [19–21]. Also, from the cross–validation tests, the jackknife test is thought the most objective approach in statistical mathematics. However, as a demonstration of practical application, predictions were also conducted for proteins in an independent dataset [7], which contains 2625 membrane proteins, of which 478 are type I transmembrane proteins, 180 type II transmembrane proteins, 1867 multipass transmembrane proteins, 14 lipid chain–anchored membrane proteins and 86 GPI–anchored membrane proteins. As a result, the success rate reaches 85%, 50%, 93%, 64% and 77% for type I membrane protein, type II membrane protein, multipass transmembrane proteins, lipid chain–anchored membrane proteins and GPI–anchored membrane proteins, respectively, and the overall success rate reaches 88%. The list of 2625 testing proteins is available upon request.

Because the extensive details for each classification (number of support vectors, the list of support vectors, the SVMlight file for prediction and the prediction results) are quite long, they are not detailed in this paper, but they are available upon request.

5 CONCLUSIONS

The above results, together with those obtained by the covariant discriminant prediction algorithm [4,7] and neural networks [6], have indicated that the types of membrane proteins are predictable with a considerable accuracy. It is anticipated that the covariant discriminant algorithm [4,7], the neural network method [6], and the SVM, if effectively complemented with each other, will become a powerful tool for predicting the types of membrane proteins. The current study has further demonstrated that the quasi–sequence–order effect as originally introduced by Chou [7] has opened a new and promising approach in dealing with sequence order effect. It has not escaped our notice that the concept of quasi–sequence–order effect as well as its mathematical framework can be used to improve the prediction quality of other protein properties as well.

6 REFERENCES

[1] B. Rost, R. Casadio, P. Fariselli, and C. Sander, Transmembrane helices predicted at 95% accuracy, *Protein Sci.* **1995**, *4*, 521–533.
[2] M. D. Resh, Myristylation and palmitylation of Src family members: the fats of the matter, *Cell* **1994**, *76*, 411–413.
[3] P. J. Casey, Protein lipidation in cell signaling, *Science* **1995**, *268*, 221–225.
[4] K. C. Chou and D.W. Elrod, Prediction of membrane protein types and subcellular locations, *Proteins Struct.*

Funct. Genet. **1999**, *34*, 137–153.
[5] A. Reinhardt and T. Hubbard, Using neural networks for prediction of the subcellular location of proteins, *Nucleic Acids Res*. **1998**, *26*, 2230–2236.
[6] Y. D. Cai, X. J. Liu and K. C. Chou, Artificial neural network model for predicting membrane protein types, *J. Biomol. Struct. Dyn*. **2001**, *18*, 607–610.
[7] K. C. Chou, Prediction of protein cellular attributes using pseudo–amino acid composition, *Proteins Struct. Funct. Genet*. **2001**, *43*, 246–255.
[8] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, 1995.
[9] V. Vapnik, *Statistical Learning Theory*, Wiley–Interscience, New York, 1998.
[10] C. H. Ding and I. Dubchak, Multi–class protein fold recognition using support vector machines and neural networks, *Bioinformatics* **2001**, *17*, 349–358.
[11] Y. D. Cai, X. J. Liu, X. B. Xu, and G. P. Zhou, Support Vector Machines for predicting protein structural class, *BMC Bioinformatics* **2001**, *2*, 3.
[12] J. R. Bock and D. A. Gough, Predicting protein–protein interactions from primary structure, *Bioinformatics* **2001**, *17*, 455–60.
[13] Y. D. Cai, X. J. Liu, X. B. Xu, and K. C. Chou, Support vector machines for prediction of protein subcellular location, *Mol. Cell Biol. Res. Commun*. **2000**, *4*, 230–233.
[14] S. J. Hua and Z. R. Sun, Support vector machine approach for protein subcellular localization prediction, *Bioinformatics* **2001**, *17*, 721–728.
[15] Y. D. Cai, X. J. Liu, X. B. Xu, and K. C. Chou, Support vector machines for prediction of protein subcellular location by incorporating quasi–sequence–order effect, *J. Cell. Biochem*. **2002**, *84*, 343–348.
[16] S. J. Hua and Z. R. Sun, A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach, *J. Mol. Biol*. **2001**, *308*, 397–407.
[17] T. Joachims, in: Making Large–Scale SVM Learning Practical. Advances in Kernel Methods – Support Vector Learning, Eds. B. Schölkopf, C. Burges, and A. Smola, MIT Press, 1999.
[18] C. Cortes and V. Vapnik, Support vector networks, *Machine Learning* **1995**, *20*, 273–293.
[19] K. C. Chou and C. T. Zhang, Prediction of protein structural classes. *Crit. Rev. Biochem. Mol. Biol*. **1995**, *30*, 275–349.
[20] Y.D. Cai, Is it a paradox or misinterpretation, *Proteins Struct. Funct. Genet*. **2001**, *43*, 36–338.
[21] Z. P. Zhou and N. Assa–Munt, Some insights into protein class prediction, *Proteins Struct. Funct. Genet*. **2001**, *44*, 57–59.