Journal of
Proteomics & Bioinformatics

Xian, J Proteomics Bioinform 2012, 5:6
DOI: 10.4172/jpb.1000225

Research Article Open Access

Prediction of Membrane Spanning β Strands in Bacterial Porins by Using Wavelet Support Vector Machine Algorithm

Guang-ming Xian*

Information Engineering and Technology Department, South China Normal University, Guangdong Foshan, China

Abstract

For accurate prediction of transmembrane β strands in bacterial porins, we proposed a wavelet support vector machine (WSVM) algorithm to predict the transmembrane β strands in bacterial porins based on the application of WSVM algorithm. The method was applied to all the five porins of known structure (three training proteins, porins *from Escherichia coli, Rhodobacter capsulatus* and *Rhodopseudomonas blastic* and two test proteins, porin from *Klebsiella pneumoniae* and *Comamonas acidovorans*). For all the five proteins the WSVM method predictived the transmembrane strands in bacterial porins to an average accuracy 84.9%, a higher predictive level than SVM (81.6%) and RNFNN (78.8%) methods. The best test result of the SVM is the precictor with wavelet kernel, which is 84.9% better than other three SVM kernel function of the Gaussian RBF kernel, Polynomial kernel as well as Linear kernel that average 81.6%, 80.3%, and 79.8%, respectively. The experimental results demonstrate the efficacy of the proposed WSVM method.

Keywords: Prediction; Membrane spanning β strands; Bacterial porin; Wavelet support vector machine; Kernel function

Introduction

Accurate and reliable prediction of protein structure and function remains a challenge [1]. Of particular importance is the prediction of membrane proteins, as, unlike soluble and fibrous proteins, membrane proteins remain poorly tractable targets for the principal experimental methods of structure determination: X-ray crystallography and multidimensional nuclear magnetic resonance (NMR) spectroscopy [2].

The membrane assembly of outer membrane proteins is more complex than that of transmembrane helical proteins owing to the intervention of many charged and polar residues in the membrane. Accordingly, the predictive accuracy of transmembrane β strands is considerably lower than that of transmembrane α helices [3].

Despite widely different functions, these proteins show a remarkable degree of structural similarity, which has led Schulz to identify 8 rules summarising β-barrel construction [4]. Of these, two are of particular importance when attempting to predict TM β-barrel topology: rule two states that both the N- and C-termini are at the periplasmic end of the barrel, restricting the strand number to even values; and rule 4, that external β-strand connections are long loops (termed L1, L2, etc.), whereas the periplasmic strand connections are generally short (T1, T2, etc.) [2].

Porins are the major component of the outer membrane of Gram negative bacteria. The crystal structures of porins have been studied in atomic detail from three different species namely, *Rhodobacter capsulatus* [5], *Escherichia coli* [6], and *Rhodopseudomonas blastica* [7], *Klebsiella pneumoniae* and *Comamonas acidovorans*. They show a common chain fold consisting of 16 anti-parallel β strands of different lengths (6-17 residues) forming a large barrel [3].

A novel type of learning machine called support vector machine (SVM) has been receiving increasing interest in areas ranging from its original application in pattern recognition to other applications such as regression estimation [8-11] due to its remarkable generalization performance. SVM was developed by Vapnik and his coworkers in 1995 [12], and it is based on the structural risk minimization (SRM) principle which seeks to minimize an upper bound of the generalization error consisting of the sum of the training error and a confidence interval [13].

SVM is a new machine learning technology that has been successfully applied in solving problems in the field of bioinformatics. A high-performance method was developed for protein secondary structure prediction based on the dual-layer support vector machine and position-specific scoring matrices (PSSMs). The SVM's performance is usually better than that of traditional machine learning approaches. The performance was further improved by combining PSSM profiles with the SVM analysis [14].

A SVM-based method was developed for predicting families and subfamilies of cytokines using dipeptide composition. The taxonomy of the cytokine superfamily with which the method complies was described in the Cytokine Family cDNA Database (dbCFC) and the dataset used in the study for training and testing was obtained from the dbCFC and Structural Classification of Proteins (SCOP). The method classified cytokines and non-cytokines with an accuracy of 92.5% by 7-fold cross-validation. The method is further able to predict seven major classes of cytokine with an overall accuracy of 94.7% [15].

A support vector machine (SVM)-based method, GPCRpred, has been developed for predicting families and subfamilies of GPCRs from the dipeptide composition of proteins. The method is further able to predict five major classes or families of GPCRs with an overall Matthew's correlation coefficient (MCC) and accuracy of 0.81 and 97.5% respectively [16].

***Corresponding author:** Guang-ming Xian, Information Engineering and Technology Department, South China Normal University, Guangdong Foshan, China, 528225, E-mail: xgm20011@126.com

Received May 12, 2012; **Accepted** June 20, 2012; **Published** June 26, 2012

Citation: Xian GM (2012) Prediction of Membrane Spanning β Strands in Bacterial Porins by Using Wavelet Support Vector Machine Algorithm. J Proteomics Bioinform 5: 135-139. doi:10.4172/jpb.1000225

For regression analysis, the non-linear ability of SVM can use kernel mapping to achieve. For the kernel mapping, the kernel function must satisfy the condition of Mercer theorem. The Gauss function is a kind of kernel function which is general used. It shows the good generalization ability. However, for our used kernel functions so far, the SVM cannot approach any curve in $L^2(R^n)$ space (quadratic continuous integral space), because the kernel function which is used now is not the complete orthonormal base. This character lead the SVM cannot approach every curve in the $L^2(R^n)$ space. Similarly, the regression SVM cannot approach every function [17].

According to the above describing, we need find a new kernel function, and this function can build a set of complete base through horizontal floating and flexing. As we know, this kind of function has already existed, and it is the wavelet functions. Based on wavelet decomposition, this paper propose a kind of allowable support vector's kernel function which is named wavelet kernel function, and we can prove that this kind of kernel function is existent. The Morlet and Mexican wavelet kernel functions are the orthonormal base of $L^2(R^n)$ space. Based on the wavelet analysis and conditions of the support vector kernel function, Morlet or Mexican wavelet kernel function for support vector regression machine (SVM) is proposed, which is a kind of approximately orthonormal function. This kernel function can simulate almost any curve in quadratic continuous integral space, thus it enhances the generalization ability of the SVM [17].

In this paper, we develop a wavelet support vector machine (WSVM) algorithm to predict the transmembrane β strands in the family of bacterial porins. A Visual C++ 6.0 program has been developed which takes the amino acid sequence as the input file and gives the predicted transmembrane β strand as output. The proposed WSVM [17] method predicts at an average accuracy level of 84.9% for all the five bacterial porins considered.

Wavelet Support Vector Machine (WSVM)

If the wavelet function $\psi(x)$ satisfied the conditions: $\psi(x) \in L^2(IR)$, and $\overline{\psi}(x) = 0$, $\overline{\psi}$ is the Fourier transform of function $\psi(x)$. The wavelet function group can be defined as

$$\overline{\psi}_{a,b}(x) = \frac{1}{\sqrt{a}}\psi\left(\frac{x-b}{a}\right), \tag{1}$$

where a is the so-called scaling parameter, b is the horizontal floating coefficient, and $\psi(x)$ is called the "mother wavelet". The parameter of translation $b(b \in R)$ and dilation a ($a > 0$), may be continuous or discrete. For the function $f(x)$, $f(x) \in L^2(R)$, The wavelet transform $f(x)$ can be defined as

$$W(a,b) = (a)^{\frac{1}{2}} \int_{-\infty}^{+\infty} f(x)\psi^*\left(\frac{x-b}{a}\right)dx, \tag{2}$$

where $\psi^*(x)$ stands for the complex conjugation of $\psi(x)$.

The wavelet transform $W\left(\{a,b\}\right)$ can be considered as functions of translation b with each scale. Eq. (2) indicates the wavelet analysis is a time-frequency analysis, or a time-scaled analysis. Different from the short time Fourier transform (STFT), the wavelet transform can be used for multi-scale analysis of a signal through dilation and translation so it can extract time-frequency features of a signal effectively.

Wavelet transform is also reversible, which provides the possibility to reconstruct the original signal. A classical inversion formula for $f(x)$ is

$$f(t) = \frac{1}{C_\psi}\int_R\int_R \frac{1}{a^2} W_f(a,b)\psi\left(\frac{t-b}{a}\right)dadb, \tag{3}$$

where $C_\psi = \int_{-\infty}^{\infty} \frac{\left|\overline{\psi}(w)\right|^2}{|w|}dw < \infty$, (4)

and $\overline{\psi}(w) = \int \psi(x)\exp(-jwx)dx.$ (5)

For the above Eq. (3), C_ψ is a constant with respect to $\psi(x)$. The theory of wavelet decomposition is to approach the function $f(x)$ by the linear combination of wavelet function group.

Wavelet kernel function

The support vector's kernel function can be described as not only the product of point, such as $k(x,x') = k\left(\langle x \cdot x' \rangle\right)$, but also the horizontal floating function, such as $k(x,x') = k\left(x - x'\right)$. In fact, if a function satisfied condition of Mercer, it is the allowable support vector kernel function.

Theorem 1: The symmetry function $k(x,x')$ is the kernel function of SVM if and only if: for all function $g \neq 0$ which satisfied the condition of $\int_{R^d} g^2(\xi)d\xi < \infty$, we need satisfy the condition as follows:

$$\int_{R^d}\int_{R^d} k(x,x')g(x)g'(x)dxdx' \geq 0 \tag{6}$$

This theorem proposed a simple method to build kernel function. For the horizontal floating function, because hardly dividing this function into two same functions, we can give the condition of horizontal floating kernel function.

Theorem 2: The horizontal floating function $k(x,x') = k\left(x - x'\right)$ is allowable support vector's kernel function if and only if the Fourier transform of $K(x)$ need satisfy the condition follows:

$$F[x](w) = (2\pi)^{-n/2}\int_{R^n} \exp(-j(w \cdot x))K(x)dx \geq 0. \tag{7}$$

If the wavelet function of one dimension is $\psi(x)$, using tensor theory, the multi-dimension wavelet function can be defined as

$$\psi_l(x) = \prod_{i=1}^{l} l\psi(x_i), x \in R^{l \times d}, \tag{8}$$

where x is a column vector with d dimensions.

We can build the horizontal floating kernel function as follows:

$$K(x,x') = \prod_{i=1}^{l} l\psi\left(\frac{x_i - x_i'}{a_i}\right), \tag{9}$$

where a_1 is the scaling parameter of wavelet, $a_1 > 0$. So far, because the wavelet kernel function must satisfy the conditions of Theorem 2, the number of wavelet kernel function which can be showed by existent functions is few. Now, we give an existent wavelet kernel function: Morlet wavelet kernel function. Morlet wavelet function is defined as follows:

$$\psi(x) = \cos(1.75x)\exp^{-\frac{x^2}{2}},\ x \in R, w_0 \in R. \tag{10}$$

Morlet wavelet kernel function is defined as

$$K(x,x') = \prod_{i=1}^{l} \cos\left(1.75 \times \frac{x_i - x_i'}{a_i}\right)\exp\left(-\frac{\left\|x_i - x_i'\right\|}{2a_i^2}\right), \tag{11}$$

where $x \in R^{l\times d}, a_i \in R,$ and this kernel function is an allowable support vector kernel function.

If we use wavelet kernel function as the support vector's kernel function, the estimation of wavelet support vector machine (WSVM) is defined as

$$f(x)=\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)\prod_{i=1}^{l}\psi\left(\frac{x^j-x_i^j}{a_i}\right)+b, b\in R. \quad (12)$$

For wavelet analysis and theory, see Krantz [18] and Liu and Di [19].

The proposed wavelet support vector machine (WSVM)

In sample set $\{(x_i,y_i), i=1,2,\ldots,\pm n\}$, $x_i \in R^n$ is used as input and y_i is used as output. And the ε intensitive loss funcion can be defined as

$$|y-f(x)|_\varepsilon=\begin{cases}0 & , \ |y-f(x)|>\varepsilon \\ |y-f(x)|-\varepsilon, & |y-f(x)|>\varepsilon\end{cases} \quad (13)$$

The optimal hyper plane of standard support vector machine can be defined as $f(x)=\langle w\cdot\Phi(x)\rangle+b$.

At the same time, combining the wavelet kernel function with support vector machine, we can build a new SVM learning algorithm that is support vector machine on wavelet kernel function (WSVM). The parameter b is taken into account confidence interval of WSVM and form the new variable w of optimal problem, then the new wavelet support vector machine whose e-insensitive tube can be reformulated as

$$\min_{w,b,\xi}\ \frac{1}{2}\|w\|^2+C\sum_{i=1}^{n}(\xi_i+\xi_i^*), \quad (14)$$

Subject to

$$|y_i-\langle w\cdot\phi(x)\rangle-b|\le\varepsilon+\xi_i \quad (15)$$

$$\xi_i^{(*)}\ge 0, \varepsilon\ge 0, \quad (16)$$

where w is a Column vector with d dimensions, $C>0$ is a penalty factor, $\xi_i^{(*)}(i=1,\ldots,l)$ are slack variables.

Problem (14) is a quadratic programming (QP) problem. By means of the Wolfe principle, wavelet kernel function technique and Karush–Kuhn–Tucker (KKT) conditions, we have the duality problem (16) of the original optimal problem (17)

$$\max_{\alpha,\alpha^*,\beta,\beta^*}(\min_{w,b,\xi}[L(w,b,\xi,\xi^*)] \quad (17)$$

and

$$\begin{aligned}L(w,b,\xi,\xi^*)&=\frac{1}{2}\|w\|^2+C\sum_{i=1}^{n}(\xi_i+\xi_i^*)-\\ &\sum_{i=1}^{n}\alpha_i(\varepsilon+\xi_i-(y_i-\langle w\cdot\phi(x)\rangle-b))\\ &-\sum_{i=1}^{n}\alpha_i^*(\varepsilon+\xi_i-(y_i-\langle w\cdot\phi(x)\rangle-b))\\ &-\sum_{i=1}^{n}(\beta_i\xi_i+\beta_i^*\xi_i^*)\end{aligned} \quad (18)$$

The original problem can be transferred to the following problem.

$$\max_{\alpha,\alpha^*,\beta,\beta^*}\left\{-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}[(\alpha_i-\alpha_i^*)\cdot(\alpha_j-\alpha_j^*)K(x_i,x_j)]-\varepsilon\sum_{i=1}^{n}(\alpha_i+\alpha_i^*)+\sum_{i=1}^{n}y_i(\alpha_i+\alpha_i^*)\right\} \quad (19)$$

Subject to

$$\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)=0, 0\le\alpha_i\le C, 0\le\alpha_i^*\le C \quad (20)$$

According to the KKT condition, we can get the following equation.

$$\begin{cases}\alpha_i[\varepsilon+\xi_i-(y_i-\langle w\cdot\phi(x_i)\rangle-b]=0\\ \alpha_i^*[\varepsilon+\xi_i^*-(y_i-\langle w\cdot\phi(x_i)\rangle-b]=0\end{cases} \quad (21)$$

Select the appropriate parameters C, and the optimal mother wavelet function which can match well the original series in some scope of scales as the kernel function of WSVM model. Then, WSVM output function is described as following:

$$f(x)=\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)\langle\Phi(x_i)\cdot\Phi(x)\rangle+b=\sum_{SV}(\alpha_i-\alpha_i^*)K(x_i,x),+b. \quad (22)$$

Database

The dataset used to derive the statistical models comprises 18 β-barrel membrane proteins of porins found in the Protein Data Bank (Table 1). All proteins share no more than 26% pairwise sequence identity.

We applied our predictive method to five different porins from (i) *Escherichia coli*, (ii) *Rhodobacter capsulatus*, (iii) *Rhodopseudomonas blastica*, (iv) *Klebsiella pneumoniae*, (v) *Comamonas acidovorans*, for which structures have been determined in atomic detail. We considered the first three proteins (*Escherichia coli*, *Rhodobacter capsulatus* and *Rhodopseudomonas blastica*) as a training set to develop the wavelet support vector machine parameter and the last two (*Klebsiella pneumoniae* and *Comamonas acidovorans*) as a test case.

Accuracy of Prediction

The accuracy of predicted segments was computed using the equation

$$\text{Accuracy}\ (\%)=[N-(N_u+N_o)]/N$$

where N, Nu and No are total number of residues, number of residues under-predicted and number of residues over-predicted in a particular protein, respectively [5].

Protein	Organism	PDB ID
OmpF	Escherichia coli	1hxx,1hxt,1hxu,1gfo,1gfq
OmpC		2jin
Porin	Rhodobacter capsulatus	2por, 3por
Porin	Rhodopseudomonas blastica	1prn, 2prn, 3prn, 4prn, 1bh3,1h6s
Ompk36	Klebsiella pneumoniae	1osm
Omp32	Comamonas acidovorans	1e54, 2fgq, 2fgr

Table 1: Dataset of 18 β-barrel membrane proteins used for this study.

Method	Protein					
	Escherichia coli	Rhodobacter capsulatus	Rhodopseudomonas blastic	Klebsiella pneumoniae	Comamonas acidovorans	Average
WSVM	85.2	84.6	86.8	84.7	83.4	84.9
SVM (RBFkernel)	84.4	82.3	82.7	81.5	77.2	81.6
SVM (Polynomial kernel)	83.2	81.4	81.3	80.1	75.5	80.3
SVM (Linear kernel)	82.1	80.2	80.7	79.6	76.4	79.8
RBFNN	81.6	78.2	80.2	79.4	74.5	78.8

Table 2: Comparison of predictive ability (%) of three other methods with the present method for bacterial porins.

Results and Discussion

We have predicted the transmembrane β strands for the all five porins and compared our WSVM results with the method of SVM with Gaussian RBF kernel function and RBF neural network (RBFNN). The experimental results are presented in (Table 2). From this table we observe that the WSVM method has the highest predictive accuracy for the protein porin from Rhodopseudomonas blastic (86.8%). Especially the proposed WSVM method predicts the transmembrane β strands of porins from *Comamonas acidovorans* with higher accuracy (83.4%) than SVM (77.2%) and RBFNN (74.5%) method. By considering all the five proteins, the proposed WSVM method predicts with higher accuracy than SVM and RBFNN methods. The proposed WSVM method predicts the transmembrane β strand for all the five proteins with >83% accuracy and the average accuracy is 84.9%.

In our study, wavelet kernel function kernel function is selected as a mapping function. The other popular kernel functions of SVM are:

Gaussian RBF kernel: $k(x_i, x) = \exp\left(-\frac{\|x - x_i\|^2}{2}\right)$

Polynomial kernel: $k(x_i, x) = (1 + x \cdot x_i)^d$

and Linear kernel: $k(x_i, x) = x_i^T x$

The best test result of the SVM is the predictor with wavelet kernel. The accuracy of SVM with wavelet kernel (84.9%) is better than other three SVM kernel function of the Gaussian RBF kernel, Polynomial kernel as well as Linear kernel that average 81.6%, 80.3%, and 79.8%, respectively. The proposed WSVM method implemented in this study can promote the accuracy of prediction of membrane spanning β strands in bacterial porin.

Conclusion

In this paper, we develop a WSVM algorithm to obtain higher accuracy in transmembrane strand prediction for the family of bacterial porins. The WSVM algorithm algorithm has been automated with a computer program written Visual C++. The proposed WSVM method predicts the membrane spanning β strands with an accuracy of >83% for all the five considered proteins (porins) individually and the average accuracy is 84.9%. These accuracy levels are superior to those of the methods of SVM and RBFNN. The results of application in prediction of membrane spanning β strands in bacterial porin show that the WSVM approach is effective and feasible, the comparison between the SVM method proposed with wavelet kernel function in this paper and other three kernel functions is also given which proves SVM with wavelet kernel function is better than SVM with other kernel functions.

References

1. Arora A, Tamm LK (2001) Biophysical approaches to membrane protein structure determination. Curr Opin Struct Biol 11: 540-547.
2. Taylor PD, Toseland CP, Attwood TK, Flower DR (2006) Beta barrel transmembrane proteins: Enhanced prediction using a Bayesian approach. Bioinformation 1: 231-233.
3. Gromiha MM, Majumdar R, Ponnuswamy PK (1997) Identification of membrane spanning beta strands in bacterial porins. Protein Eng 10: 497-500.
4. Schulz GE (2000) beta-Barrel membrane proteins. Curr Opin Struct Biol 10: 443-447.
5. Weiss MS, Schulz GE (1992) Structure of porin refined at 1.8 A resolution. J Mol Biol 227: 493-509.
6. Cowan SW, Schirmer T, Rummel G, Steiert M, Ghosh R, et al. (1992) Crystal structures explain functional properties of two E. coli porins. Nature 358: 727-733.
7. Kreusch A, Schulz GE (1994) Refined structure of the porin from Rhodopseudomonas blastica. Comparison with the porin from Rhodobacter capsulatus. J Mol Biol 243: 891-905.
8. Muller R, Smola JA, Scholkopf B (1997) Prediction time series with support vector machine in. Proc Int. Conf Artificial Neural Networks 999.
9. Mukherjee S, Osuna E, Girosi F (1997) Nonlinear prediction of chaotic time series using support vector machine. IEEE Signal Proceesing Society Workshop on Neural Networks for Signal Processing VII (NNSP'97), Amelia Island, FL 511-520.
10. Vapnik VN, Golowich SE, Smola AJ (1996) Support vector method for function approximation, regression estimation and signal processing. Advances in Neural Information Processing Systems.
11. Muller KR, Smola JA, Ratsch G, Scholkopf B, Kohlmorgen J (1999) Prediction time series with support vector Machine. Advances in Kernel Methods, MIT Press, London, U.K.
12. Vapnik VN (1995) The Nature of Statistical Learning Theory. Statistics for Engineering and Information Science, 2nd Edition, Springer-Verlag, New York.
13. Cao LJ, Tay FH (2003) Support Vector Machine with Adaptive Parameters in Financial Time Series Forecasting. IEEE Trans Neural Netw 14: 1506-1518.
14. Guo J, Chen H, Sun Z, Lin Y (2004) A novel method for protein secondary structure prediction using dual-layer SVM and profiles. Proteins 54: 738-743.
15. Huang N, Chen H, Sun Z (2005) CTKPred: an SVM-based method for the prediction and classification of the cytokine superfamily. Protein Eng Des Sel 18: 365-368.

16. Bhasin M, Raghava GP (2004) GPCRpred: an SVM-based method for prediction of families and subfamilies of G-protein coupled receptors. Nucleic Acids Res 32: W383-W389.

17. Qi Wu (2009) The forecasting model based on wavelet m-support vector machine. Expert Systems with Applications 36: 7604-7610.

18. Krantz SG (1994) Wavelet: Mathematics and application. Boca Raton, FL: CRC.

19. Liu GZ, Di SL (1992) Wavelet Analysis and Its Applications. Xi'an, Xidian University Press House, China 23-28.