PROTEINS: Structure, Function, and Bioinformatics 56:11–18 (2004)

SHORT COMMUNICATION

Prediction of Transmembrane Regions of β-Barrel Proteins Using ANN- and SVM-Based Methods

Navjyot K. Natt, Harpreet Kaur, and G. P. S. Raghava*
Institute of Microbial Technology, Chandigarh, India

ABSTRACT **This article describes a method developed for predicting transmembrane β-barrel regions in membrane proteins using machine learning techniques: artificial neural network (ANN) and support vector machine (SVM). The ANN used in this study is a feed-forward neural network with a standard back-propagation training algorithm. The accuracy of the ANN-based method improved significantly, from 70.4% to 80.5%, when evolutionary information was added to a single sequence as a multiple sequence alignment obtained from PSI-BLAST. We have also developed an SVM-based method using a primary sequence as input and achieved an accuracy of 77.4%. The SVM model was modified by adding 36 physicochemical parameters to the amino acid sequence information. Finally, ANN- and SVM-based methods were combined to utilize the full potential of both techniques. The accuracy and Matthews correlation coefficient (MCC) value of SVM, ANN, and combined method are 78.5%, 80.5%, and 81.8%, and 0.55, 0.63, and 0.64, respectively. These methods were trained and tested on a nonredundant data set of 16 proteins, and performance was evaluated using "leave one out cross-validation" (LOOCV). Based on this study, we have developed a Web server, TBBPred, for predicting transmembrane β-barrel regions in proteins (available at http://www.imtech.res.in/raghava/tbbpred). Proteins 2004;56:11–18.** © 2004 Wiley-Liss, Inc.

Key words: β-barrels; transmembrane proteins; artificial neural networks; multiple sequence alignment; support vector machine; physicochemical parameters; LOOCV

INTRODUCTION

Integral membrane proteins play a central role in cellular metabolism. Approximately 20% of genes may encode for membrane proteins.[1,2] Among them, plasma membrane proteins consist of mostly transmembrane α-helices and outer membrane proteins consisting of β-barrels. Thus, compared to soluble cytoplasmic globular proteins having a large number of possible structural folds, integral membrane proteins exist in more abundant transmembrane helices and rather less encountered β-barrel proteins.[2,3] Both types show high neighborhood co-relation limiting the total number of different topologies. The latter have been known in the outer membrane of bacteria, chloroplasts, and mitochondria. But none of the chloroplast and mitochondrial proteins have yet been structurally proven. Functions of β-barrel membrane proteins are more diverse than anticipated earlier, when they were considered simple passive pores used for transport across bacterial membranes.[4,5] Their functions are as diverse as active ion transporters for nutrient uptake, membrane anchors, membrane-bound enzymes, and also for defense against pathogenic proteins. It is now evident that different barrel sizes are associated with different functions.

Presently, known sizes range from small, 8-stranded to large, 22-stranded β-barrels existing either as monomers or oligomers. The smallest monomeric barrels form inverse micelles and work as enzymes or bind to macromolecules, or are involved in pathogenicity (e.g., OmpT, OmpLA).[6,7] The medium-range barrels, which include trimeric porins of gram-negative proteins, form more or less specific pores for nutrient uptake (e.g., OmpX), while the largest barrels occur in active Fe+2 transporters (e.g., FhuA, FepA).[8,9]

All β-barrels contain meandering, even-numbered antiparallel sheets, whose topologies are defined by their strand number and shear number (measure of inclination angle of β-strand against the axis). A set of 10 construction rules have been outlined by Schulz.[5] The membrane assembly of outer membrane proteins is more complex than that of transmembrane helical proteins, owing to intervention of many charged and polar residues in the membrane.[10] The simplest approach of looking for alternating polar and nonpolar residues at the inside and outside of membrane, which proved successful for transmembrane helix prediction,[11,12] is thus not applicable for β-barrel proteins. Also, the development of a three-dimensional (3D) model for transmembrane β-barrel proteins has not

*Correspondence to: G. P. S. Raghava, Bioinformatics Centre, Institute of Microbial Technology, Sector 39A, Chandigarh, India. E-mail: raghava@imtech.res.in

Received 25 September 2003; Accepted 22 December 2003

Published online 7 May 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20092

been undertaken extensively due to low sequence identity among these proteins even in the membrane spanning regions; thus, only a single model is available to date.[13] However, a number of methods are available for transmembrane helix prediction, such as DAS, PHDhtm, SOSUI, PRED-TMR, TMHMM, and TopPred2,[14] in contrast to only a few methods available for transmembrane β-barrel prediction.[15–17]

A rules-based approach has been successfully applied for the prediction of transmembrane β-barrels in porins by Gromiha et al.[18] A similar approach based on the analysis of 3D structures of 6 β-barrel proteins including parameters such as secondary structure, hydropathy, and amphipathicity, was undertaken by Zhai et al. in 2002.[17] A neural network-based predictor was developed that is especially suited to predict the topography of the β-barrel transmembrane proteins.[15] A hidden Markov Model (HMM)-based method, trained on evolutionary information, was developed, which predicted not only the transmembrane β-barrel regions but also discriminated between outer membrane, transmembrane helical proteins, and cytoplasmic globular proteins.[16] However, to the best of our knowledge, there is not a single public domain Web server for predicting the β-barrel region in proteins.

In this study, a systematic attempt has been made to develop the method for predicting transmembrane β-barrel regions in proteins using the machine learning techniques and larger data set of proteins. The machine learning techniques in this study include the artificial neural network (ANN) and the support vector machine (SVM). The ANN used is a feed-forward neural network with a standard back-propagation training algorithm. It has been shown in the past that evolutionary information in the form of multiple sequence alignment or profile improves the performance of secondary structure prediction methods.[19–21] Thus, we have used multiple sequence alignment obtained from PSI-BLAST[22] output as input for ANN instead of single sequence. We have also implemented another powerful and commonly used classifier SVM for predicting β-barrels. SVM, a class of statistical learning algorithm, was first explained by Vapnik.[23] Since then, they have become extremely popular and are being enthusiastically implemented in the field of computational biology for protein fold recognition, Major Histocompatibility Complex (MHC) binding peptide prediction, microarray analysis, and many other classification problems. In our present work, two kinds of SVM models were developed — one based only on the amino acid sequence information and a second based on amino acid information and 36 physicochemical parameters. The parameters used for developing this model are available with Protein Sequence Analysis (PSA) server at http://www.imtech.res.in/raghava/psa. The PSA Web server requires only protein sequence. The server has data for all 36 of the above-mentioned physicochemical parameters for all 20 amino acids collected from the literature. The major classes of parameters used were solvent accessibility, hydrophobicity, hydrophilicity, flexibilty, charge, volume, polarity, concentration of neighboring aromatic residues, and propensities of all 20 amino acids for α-helices, β-sheets, and turns. We have exploited the ability of SVM to work on a very large number of feature vectors by using information in the form of 36 physicochemical parameters, including parameters associated with various propensities of all 20 amino acids, along with sequence information. In the earlier studies, conformational parameter information, along with hydrophobicity profiles of amino acids, was used by Gromiha et al.,[18] but only for porins as a rules-based approach. This set of features was motivated by the distinct hydrophobicity profile of transmembrane β-strands, flexibilty of the interconnecting loops that are part of barrels, and finally, the solvent accessibilty and polarity profiles that mark the cytoplasmic and transmembrane regions of the β-barrels.

The overall per residue accuracy is 78.5% and 80.5% for SVM- and ANN-based methods, respectively. Finally, a combination of the two learning techniques has resulted in a better accuracy measure of 81.8%, which is comparable to the best available method.[15] The accuracy of ANN and SVM for discriminating between β-barrel and non-β-barrel proteins is 88.9% and 92.3%, respectively. A Web server based on this method is available as Tbbpred at http://www.imtech.res.in/raghava/tbbpred.

SYSTEM AND METHODS

Data Set

The data set used for training and testing consisted of 16 nonredundant β-barrel membrane proteins with less than 30% sequence similarity, and whose 3D structure information is available in the Protein Data Bank (PDB). The number of β-strands in the barrels varied from 8 to 22. Proteins and their detailed description, including PDB codes, are shown in Table I.

To discriminate between β-barrel and non-β-barrel proteins, a data set of 116 proteins was created, which consisted of 100 globular proteins having less than 25% sequence identity and 16 β-barrel proteins (used in this study).

Implementation of the Neural Network Predictor

Neural network predictor was implemented using the Stuttgart neural network simulator (SNNS). A feed-forward neural network with back-propagation algorithm was used to discriminate between membrane β-strand and non-β-strand regions. The network consisted of one hidden layer of 5 nodes and a single output node. A window size of 9 was used. Evolutionary profiles were given as input, as derived from PSI-BLAST[22] with threshold 0.001 to search against the **nr** database available at http://ncbi.nlm.nih.gov/blast/.

Sequence-Based SVM Model

Feature representation. In the case of sequence-based model, for each amino acid of the data set protein, feature vectors were assembled from binary encoded representations of each individual amino acid only. A window size of 9 was used, and each amino acid was represented by 21 units. The SVM model was trained and tested on the data

TABLE I. Details of Proteins Included in the Data Set, Where Number Refers to the Number of Strands in a Barrel

PDB code	Organism	Name & Number	Authors
1A0S	*Salmonella typhimurium*	Maltoporin—18	Forst et al.[27]
2MPR	*Salmonella typhimurium*	Maltoporin—18	Meyer et al.[28]
1Af6	*Echerichia coli*	Maltoporin—18	Wang et al.[29]
1BXW	*Echerichia coli*	OmpA—8	Pautsch and Schulz[30]
1QJ8	*Echerichia coli*	OmpX—8	Vogt and Schulz[31]
1E54	*Commanus acidovorans*	Porins—16	Zeth et al.[7]
1PRN	*Rhodopseudomonas blastica*	Porins—16	Kreutsch and Schulz[32]
2POR	*Rhodobacter capsulatus*	Porins—16	Weiss and Schulz[33]
2OMF	*Echerichia coli*	Porins—16	Cowan et al.[34]
1FCP	*Echerichia coli*	Fe+2 Transporter—22	Fergusson et al.[9]
1QKC	*Echerichia coli*	Fe+2 Transporter—22	Fergusson et al.[9]
1FEP	*Echerichia coli*	Fe+2 Transporter—22	Buchman et al.[8]
1I78	*Echerichia coli*	OmpT—12	Krammer et al.[35]
1QD5	*Echerichia coli*	Phospholipase A—22	Snijder et al.[36]
1PHO	*Echerichia coli*	Phosphoporin—18	Cowan et al.[37]
1K24	*Neisseria meningitidis*	OpcA—10	Prince and Actin[38]

set of 16 proteins and evaluated using "leave one out cross-validation" (LOOCV).

Physicochemical Parameter-Based Model

Feature Representation. For constructing this model, the real encoded values representing various physicochemical features were added to the amino acid sequence information (Table II). A window size of 9 was used. Each amino acid was represented by 21 binary encoded units plus 36 real encoded units. The parameter values were averaged to fit in the range of 0–1, as amino acid information was binary encoded.

$$\mathrm{AvgParm} = \frac{\mathrm{Parm} - \mathrm{Min}_{\mathrm{val}}}{\mathrm{Max}_{\mathrm{val}} - \mathrm{Min}_{\mathrm{val}}}. \quad (1)$$

Here, AvgParm represents any averaged physicochemical parameter and Parm, $\mathrm{Min}_{\mathrm{val}}$, and $\mathrm{Max}_{\mathrm{val}}$ represent the actual value, maximum value, and minimum value, respectively, of the parameter under consideration.

Implementation. SVM learning was implemented using SVMlight (Joachims[24]), available at http://www-ai.cs.uni-dortmund.de/software/svm_light. This package enables the user to define the number of parameters and also the choice of various kernels as linear, polynomial, radial basis function, sigmoid, or any user-defined kernel. In this study, the regression mode of SVM was used to model the transmembrane β-strand regions of the training set.

Assuming that we have number of proteins $x_i \in R^d (I = 1,2\ldots N)$ with corresponding target values $y_i \in \{\langle \text{target value} \rangle\}$, the x_i corresponds to the sequence of amino acids representing a protein presented to SVM for learning. Here, the target value is either +1, representing an amino acid in the transmembrane β-strand, or −1, representing a residue in conformation other than transmembrane β-barrel. SVM maps the input vectors x_i into high-dimensional space, where the error is minimal on the training set. The decision function implemented by SVM can be written as

$$F(x) = \mathrm{sign}\left(\sum_{i=1}^{N} y_i \alpha_i K(x_i x_j - b)\right). \quad (2)$$

The value of α_i is given by the task of quadratic programming, thus maximizing the subject to $0 \le \alpha_i \le C$. C is the regulatory parameter that controls the trade-off between the margin and the training error and b is the threshold for defining the hyperplane.

Training and testing sets were developed from the data set of 16 proteins. In this work, various parameters related to SVM learning were chosen and optimized after spending lots of computational time. The kernel chosen was Radial Basis Function (RBF) with regression mode. While the choice of RBF width is at least guided by the heuristic, there is no hint available on how to choose the error weight C. Choosing the kernel type is analogous to choosing architecture for artificial neural networks. The C parameter that controls the error–margin trade-off was set to 30, and parameter γ-g was set to 0.08; J—the cost factor by which the training errors on positive examples outweigh errors on negative examples—was set to 0.1. Learning was carried out using other kernels—linear, polynomial, and sigmoid also—but the best results were obtained with the RBF kernel:

$$K_{\mathrm{RBF}}(x_1,x_2) = \exp\left(\frac{-\|x_1 - x_2\|^2}{2\sigma^2}\right). \quad (3)$$

Combination of SVM and ANN Methods

In this study, finally, we have combined both SVM and ANN methods (for details, see architecture in Fig. 1). In the case of SVM, output of a residue was in the range of −1.5 to 1.5 compared to 0 to 1 in ANN. We have normalized the SVM score in order to make it in the range of 0 to 1 by adding 1.5 to the SVM score and dividing by 3. The final per residue score was calculated by taking the average of two scores (ANN score and normalized SVM score).

TABLE II. All 36 Properties, Which Include Various Physicochemical Properties and Conformational Propensities for all 20 Amino Acids

Hydrophobicity	Method
	Avg. surrounding hydrophobicity—Manvalin et al.[39]
	Hydrophobic index—Ponnuswamy et al.[40]
	Hydrophobicity in folded form—Ponnuswamy et al.[40]
	Hydrophobic gain—Ponnuswamy et al.[40]
	Surr. hydrophobicity in α-helix—Ponnuswamy et al.[40]
	Surr. hydrophobicity in β-sheet—Ponnuswamy et al.[40]
	Surr. hydrophobicity in β-turn—Ponnuswamy et al.[40]
	Hydrophobicity - Eisenberg[41]
Hydrophilicity	
	Hydrophilicity—Hopp and Woods[42]
	Hydropathy—Kyte and Dolittle[43]
	Hydrophilicity from HPLC—Parker et al.[44]
	Hydrophilicity—Jones[45]
Solvent accesibility	
	Percentage of buried residues—Janin et al.[46]
	Percentage of exposed residues—Janin et al.[46]
	Accesibility reduction ratio—Ponnuswamy et al.[41]
	Avg. number of surrounding residues—Ponnuswamy et al.[41]
Flexibity	
	Flexibility—Bhaskaran and Ponnuswamy[47]
	Flexibility for no rigid neighbors—Bhaskaran et al.[48]
	Flexibility for 1 rigid neighbor—Bhaskaran et al.[48]
	Flexibility for 2 rigid neighbors—Bhaskaran et al.[48]
	Local flexibility—Ragone et al.[49]
Free energy transfer to surface	Bull and Brease[50]
Polarity	Ponnuswamy et al.[40]
Volume	Chothia et al.[51]
Normalized frequency of β-sheet, α-helix, and reverse turn	Levith et al.[57]
Charge	—
Local conc. Ar residues	—
OMH	Barrel and Bankier[57]
PEST	Signature of rapidly degrading proteins

HPLC, ____; OMH, ____; PEST, ____.

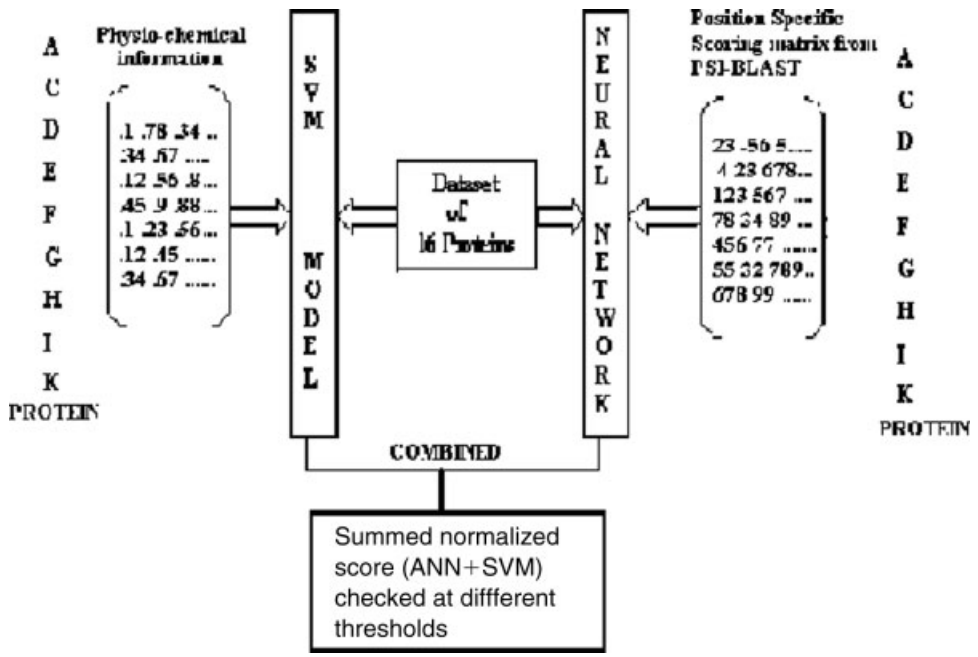

Fig. 1. Pictorial representation of the prediction method.

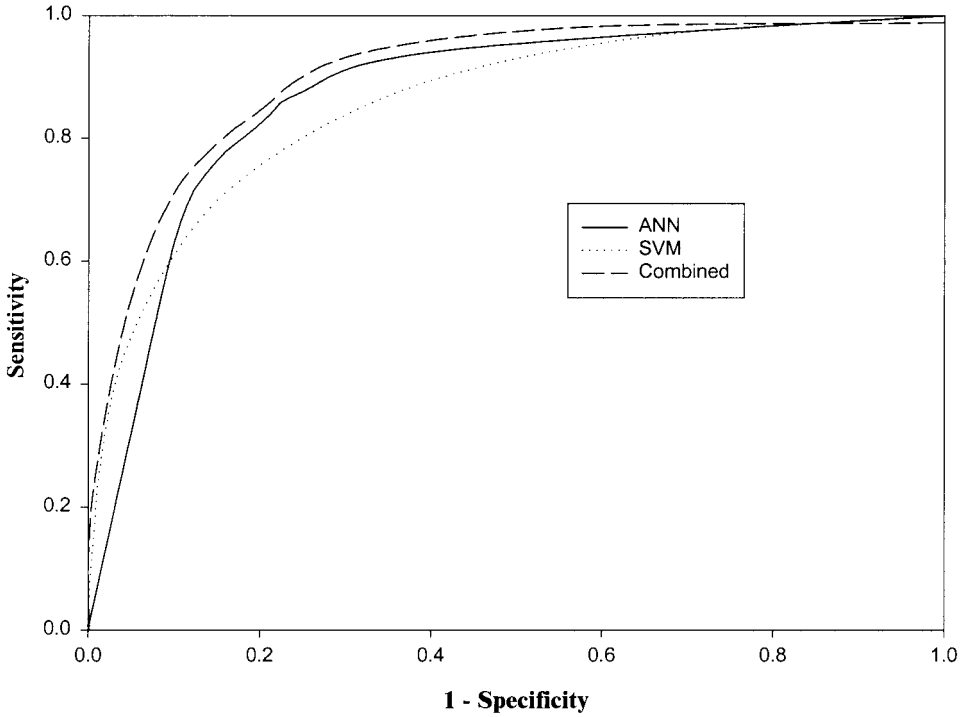

Fig. 2. ROC plot for ANN, SVM, and the combined prediction method.

Evaluating the Model/Network

The performance of any prediction algorithm is often checked by the cross-validation or jack-knife tests.[25] The prediction accuracy of both the trained neural network and SVM model was tested using LOOCV. This is considered to be one of the most rigorous testing procedures, wherein the entire data set of n proteins is divided into n subsets. The classifier was trained on n-1 proteins and tested on the nth protein. The entire process was repeated n times using each subset as the test set and the rest as a training set. The results of the test sets were combined to get an overall estimate of prediction accuracy and finally checked at different thresholds.

Scoring the Prediction

The prediction results from both SVM and ANN were evaluated using the following statistical measures:

1. Accuracy of all the three methods—neural network, SVM model, and combined (Fig. 2)—was calculated as follows:

$$Q_{\text{Acc}} = \frac{\text{TOTc}}{\text{Total}}, \tag{4}$$

where TOTc is the total number of correct predictions (includes both true postives and true negatives) and Total is the total number of predictions made.

2. The Matthews correlation coefficient (MCC) is defined as

$$\text{MCC} = \frac{(P \times N) - (O \times U)}{\sqrt{(P + U) \times (P + O) \times (N + U) \times (N + O)}}, \tag{5}$$

where P and N refer to correct β and non-β predictions, and O and U refer to over- and underpredictions, respectively.

3. Sensitivity (Q_{sens}), specificity (Q_{spec}), and non-predicted value (NPV) of the prediction methods are defined as

$$Q_{\text{sens}} = \frac{TP}{TP + FN}$$

$$Q_{\text{spec}} = \frac{TN}{TN + FP} \tag{6}$$

$$\text{NPV} = \frac{TN}{FN + TN},$$

where TP, FP, TN, and FN refer to true positives, false positives, true negatives, and false negatives, respectively.

4. The Qok index was calculated as the percentage of proteins with correctly located β-strands with tolerance $T = 4$. This was calculated as described by Rost et al.[14] Qok = number of correctly predicted proteins/total number of proteins.

RESULTS AND DISCUSSION

Prediction of β-barrel regions from the primary sequence of transmembrane β-barrel proteins has not been

TABLE III. Statistical Evaluation of SVM Model and Neural Network Predictor

Method	Accuracy	Senstivity	Specificity	NPV	MCC	Qok
SNNS–Sequence	70.38	0.67	0.75	0.67	0.41	—
SNNS–PSI-BLAST	80.52	0.86	0.78	0.86	0.63	0.4375 (7/16)
SVM (seq. only)	77.38	0.72	0.83	0.80	0.54	—
SVM (Seq. + parm.)	78.45	0.70	0.85	0.80	0.55	0.625 (10/16)
SNNS + SVM	81.83	0.82	0.83	0.86	0.64	0.5625 (9/16)

*The values in parenthesis are the number of proteins predicted correctly.

extensively undertaken due to the limited number of structures available, very low sequence identity even in the transmembrane regions, and the hydrophobicity profile (characteristic of transmembrane proteins), which does not follow a set pattern. Also, among the available methods, none is available as a Web server.

Here, we have explored the learning potentials of machine-learning techniques, namely, ANN and SVM. The two main inspirations behind the present work were (1) to make available a Web-based tool for the prediction of transmembrane β-barrels, and (2) to exploit the learning potential of SVM, especially in the case of nonlinear mappings, and also utilize the characteristic nature of amino acids present in the various regions of the transmembrane β-strands.

The results obtained by the two methods provided different insights. The ANN-based method, which utilized both sequence information and evolutionary profiles of the data set proteins, showed similar results to those obtained in the first attempt by Casadio et al.[16] However, in the present case with an updated data set of 16 proteins, the prediction accuracy is 80.5% compared to 78.0% reported earlier with evolutionary profile and 70.38% as compared to 69.3% with just sequence information (Table III). This is probably due to the fact that a larger number of proteins was used for training and testing in the present study. Another probable reason could be the more extensive learning, which we undertook using lots of variation in the learning parameters of the neural network. This highlights the ever-contested feature of ANN being a "black box" method due to the lack of interpretation of the weights achieved in the optmization process and the fact that solution is not unique, because different weights and weight patterns can lead to the same or better prediction outcome. On the other hand, SVM training always seeks globally optimized solution and has the potential to deal with a large number of feature vectors. Another advantage of SVM over neural networks is its transparency. Prediction accuracy achieved by an SVM-trained model using LOOCV is 78.45%, which is comparable to that achieved by ANN-based learning.

Earlier work by Schulz et al.[26] has shown the characteristics of various amino acids in the transmembrane β-strands and their relevance. The hydrophobicity and hydrophilicity profiles have always been considered characteristic of transmembrane proteins. These features were exploited to develop an SVM model for the prediction of residues in transmembrane β-barrel regions. As shown in Table III, the combined predictions resulted in an increase in prediction accuracy (81.8%), which is comparable to the best available method.[15] The fundamental reason that can be speculated from the results is the basic difference in the information used for the training SVM model and neural network predictor; whereas the SVM-based model worked on sequence information and physicochemical parameters, the neural network predictor leared from evolutionary profiles obtained from PSI-BLAST. Because of this difference in the learning techniques, one method learned what the other one failed to learn. This is the first report of the development of an SVM-based method for prediction of transmembrane β-barrel regions. The methods described in this study are threshold dependent; thus, their accuracy also depends on threshold. We have selected the threshold at which sensitivity and specificity values are nearly the same. The performance of ANN, SVM, and the combined approach is presented as a receiver operating characteristic (ROC) plot, which further demonstrates the better quality of prediction by the combined method. One of the most important measures of accuracy in the case of membrane proteins is Qok. This measure is also checked for the two methods (see Table III).

Discriminating Power

The method described in this study was developed with the aim of predicting β-barrel regions in transmembrane β-barrel proteins. Thus, one of the major problems is to discriminate between globular and transmembrane β-barrel proteins. We have made an attempt to use this method for scanning/classifying the proteins. A data set consisting of all 16 β-barrel proteins used in this study and 100 globular proteins (having less than 25% sequence identity) picked up randomly, was used. The SVM- and ANN-based methods were used to predict β-barrel regions in all 116 proteins in the data set. We know that β-barrel proteins consist of a large number of β-barrel regions; for example, all 16 β-barrel proteins used in this study have minimum of 8 β-strands, where the minimum length of each strand is 5 residues. In order to classify a protein as β-barrel or non-β-barrel, we have chosen the criterion of a minimum of 8 strands. The discriminative capability of SVM and ANN methods is shown in Table IV, which indicates that they were able to classify two classes (globular and transmembrane β-barrel) of proteins with respective accuracies of 92.3% and 88.8%, respectively, at a minimum length of 7 residues. Thus, the method can also be used to classify proteins.

TABLE IV. Discrimination Power of the SVM Model and Neural Network Based on Data Set of 116 Proteins (16 Transmembrane Proteins + 100 Globular Proteins)

	Strand Length			
	5	6	7	8
ANN Strand No.				
4	0.376	0.478	0.589	0.769
6	0.521	0.649	0.786	0.871
8	0.684	0.769	0.889	0.914
SVM Strand No.				
4	0.171	0.188	0.385	0.786
6	0.188	0.325	0.752	0.923
8	0.299	0.589	0.923	0.923

TbbPred Server

Based on this work, the Web server TbbPred has been developed to predict the transmembrane β-barrel regions from the primary sequence of a protein. The server implements both neural network predictor and the SVM model, wherein the user can choose either the neural network or SVM, or a combined approach for prediction. Further, the method also predicts whether the query protein is a β-barrel membrane protein or not. The server is available at http://www.imtech.res.in/raghava/tbbpred.

REFERENCES

1. Wallin E, Von Heijne G. Genome wide analysis of integral membrane proteins eubacterial, archaean and eukaryotic organisms. Protein Sci 1998;7:1029–1038.
2. Jones DT. Do transmembrane protein superfolds exist? FEBS Lett 1998;423:281–285.
3. Cowan SW, Rosenbusch JP. Folding pattern diversity of integral membrane proteins. Science 1994;264:914–916.
4. Schulz GE. Porins: general to specific, native to engineered passive pores. Curr Opin Struct Biol 1996;10:485–490.
5. Schulz GE. β-Barrel membrane proteins. Curr Opin Struct Biol 2000;10:443–447.
6. Jenkins JA, Karlson R, Konig N, Paupit RA, Jasonius JN, Rizkallah PJ, Rosenbusch JP, Rummel G, Schirmer T. The structure of OmpF porin in a tetragonal crystal form. Structure 1995;3:1041–1050.
7. Zeth K, Deiderichs K, Welte W, Engelhardt H. Crystal structure of Omp32 , the anion selective porin from *Comamonas acidovorans* in complex with a periplasmic peptide at 2.1 Å resolution. Structure 2000;8:981–992.
8. Buchman SK, Smith BS, Venkatramani L, Xia D Esser L, Palnitkar M, Chakraborty R, VanderHalm D, Deisenhofer J. Crystal structure of the outer membrane active transporter FepA from *Escherchia coli*. Nat Struct Biol 1999;6:56–63.
9. Cowan SW, Garavito RM, Jasonius JN, Fergusson AD, Hofmann E, Coulton JW, Diedericks K, Welte W. Siderophore-mediated iron transport: crystal structure of FhuA with bound lipopolysacharride. Science 1998;282:2215–2220.
10. Ferguson AD, Braun V, Fielder HP, Coulton JW, Diederichs K, Welte W. Crystal structure of the antibiotic albomycin in complex with the outer membrane protein FhuA. Protein Sci 2000;9:956–9639.
11. Tusnady GE, Simon I. Principles governing amino acid composition of integral membrane proteins: application to topology prediction. J Mol Biol 1998;283:489–506.
12. Rost B, Farseilli P, Sander C. Transmembrane helices predicted at 95% accuracy Protein Sci 1995;4:521–533.
13. Rost B, Farseilli P, Casadio R. Topology prediction for helical transmembrane proteins at 86% accuracy. Protein Sci 1996;5:1704–1718.
14. Casadio R, Jacoboni I, Messina A, Pinto V. A 3-D model of the voltage-dependent anion channel (VDAC). FEBS Lett 2002;520:1–7.
15. Rost B, Chen CP, Kernytsky A. Transmembrane helix predictions revisited. Prot Sci 2002;11:2774–2791.
16. Jacoboni I, Martelli PL, Farselli P, De Pinto V, Casadio R. Prediction of the transmembrane regions of the β-barrel membrane proteins with neural network-based predictor. Protein Sci 2001;10:779–787.
17. Martelli RL, Fariselli P, Krogh A, Casadio R. A sequence profile based HMM for predicting and discriminating β-barrel membrane proteins. Bioinformatics 2002;18:S46–S53.
18. Zhai Y, Saier MH Jr. The β-barrel finder (BBF) program, allowing identification of outer membrane β-barrel proteins encoded within prokaryotic genomes. Prot Sci 2002;11:2196–2207.
19. Gromiha MM, Majumdar R, Ponnuswamy PK. Identification of membrane spanning β-strands in bacterial porins. Protein Eng 1997;5:497–500.
20. Cuff JA, Barton GJ. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins 2000;3:502–511.
21. Kaur H, Raghava GP. A neural-network based method for prediction of gamma-turns in proteins from multiple sequence alignment. Protein Sci 2003;5:923–929.
22. Kaur H, Raghava GP. Prediction of beta-turns in proteins from multiple alignment using neural network. Protein Sci 2003;3:627–634.
23. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acid Res 1997;25:3389–3402.
24. Vapnik V, Chervonenkis A. Theory of Pattern recognition. (In Russian) Moscow: Nauka, 1974.
25. Joachims T. Text categorization with support vector machines: learning with many relevant features. Proceedings of the European conference on machine learning. Springer, Berlin.
26. Chou KC, Zhang CT. Prediction of protein structural classes. Crit Rev Biochem Mol Biol 1995;30:275–349.
27. Schulz GE. The structure of bacterial outer membrane proteins. Biochimica et Biophysica Acta 2002;1565:308–317.
28. Sansom SP, Ulmschineider MB. Amino acid distributions in integral membrane protein structures. Biochimica et Biophysica Acta 2001;1512:1–14.
29. Forst D, Welte W, Wacker T, Diederichs K. Structure of the sucrose specific porin ScrY from *Salmonella typhimurium* and its complex with sucrose. Nat Struct Biol 1998;5:37–46.
30. Meyer JE, Hofnung M., Schulz GE. Structure of maltoporin from *Salmonella typhimurium* ligated with nitrophenyl maltotrioside. J Mol Biol 1997;226:761–775.
31. Wang YF, Dutzler R, Rizakallah PJ, Rosenbusch JP, Schirmer T. Channel specificity: Structural basis for sugar discrimination and differential flux rates in maltoporin. J Mol Biol 1997;272:56–63.
32. Pautsch A, Schulz GE. Structure of the outer membrane protein A transmembrane domain. Nat Str Biol 1998;5:1013–1017.
33. Vogt J, Schulz GE. The structure of the outer membrane protein OmpX from *Escherichia coli* reveals mechanisms of virulence. Structure 1999;7:1301–1309.
34. Kreusch A, Schulz GE. Refined structure of the porin from *Rhodobacter blastica*: comparison with the porin from *Rhodobacter capsulatus*. J Mol Biol 1994;243:891–905.
35. Weiss MS, Schulz GE. Structure of porin refined at 1.8 Å resolution. J Mol Biol 1992;227:493–509.
36. Cowan SW. Refined structure of OMF porin from E. Coli at 2.3 Å resolution. (unpublished).
37. Vandeputte-Rutten L, Kramer RA, Kroon J, Dekker N, Egmond MR, Gros P. Crystal structure of outer membrane protease OmpT from Escherichia coli suggests a novel catalytic site. EMBO 2001;20:5033–5039.
38. Snijder HJ, Ubarretxena-Belandia I, Blaauw ML. Structural evidence for dimerization-regulated activation of an integral membrane phospholipase. Nature 1999;401:717–721.
39. Cowan SW, Schirmer T, Rummel G, Steirt M, Ghosh R, Paupit RA, Jasonius JN, Rosenbusch JP. Crystal structures explain functional properties of two *E. coli* porins. Nature 1992;358:727–733.
40. Prince SM, Actman M, Derrick JP. Crystal Structure of OpcA, integral membrane protein from *Neisseria meningitis.* Proc Natl Acad Sci USA 2002;99:3417–3421.

41. Manvalan P, Ponnuswamy PK. Hydrophobic character of amino acids in globular proteins. Nature 1978;275:673–674.
42. Ponnuswamy PK, Prabhakaran M, Manvalan P. Hydrophobic packing and spatial arrangement of amino acids in globular proteins. Biochimica et Biophysica Acta 1980;623:301–316.
43. Eisenberg D, Schwarz E, Komarony M, Wall R. Analysis of membrane and surface protein sequences with hydrophobic moment plot. J Mol Biol 1984;179:125–142.
44. Hopp TP, Woods KR. Prediction of protein antigenic determinants from amino acid sequences. Proc Natl Acad Sci USA 1981;78:3824–3828.
45. Kyte J, Doolittle RF. A simplest method for displaying the hydropathic character of the proteins. J Mol Biol 1982;157:105–132.
46. Parker JM, Guo D, Hodges RS. New hydrohilicity scale derived from high performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and X-ray derived accessible sites. Biochemistry 1986;25:5425–5432.
47. Jones DD. Amino acid properties and side chain orientation in proteins: a cross-correlation approach. J Theo Biol 1975;50:167–183.
48. Janin J, Wodak S. Conformation of amino acid side chains in proteins. J Mol Biol 1978;125:357–386.
49. Bhaskaran, Ponnuswamy PK. Possible flexibilities of amino acid residues in globular proteins. Int J Peptide Protein Res 1988;32:241–255.
50. Karplus PA, Schulz GE. Naturwissens-chaften 1985;72:212–219.
51. Ragone R, Facchiano F, Facchiano AM, Colonna G. Flexibility plot of proteins. Protein Engg 1989;2:497–504.
52. Bull HB, Breese K. Surface tension of amino acid solution—a hydrophobicity scale of the amino acid residues. Archi Bioch Biophy 1974;161:665–670.
53. Chothia C. Principles that determine the structure of proteins. Annu Rev Biochem 1984;53:537–572.
54. Levitt M. Conformational Preferences of amino acids in globular proteins. Biochemistry 1978;17:4277–4285.
55. Barrel BG, Bankier AT, Drouin J. A different genetic code for human mitochondrial genome. Nature 1979;282:189–194.