SVMtm: Support Vector Machines to Predict Transmembrane Segments

ZHENG YUAN,[1] JOHN S. MATTICK,[2] ROHAN D. TEASDALE[1]
[1]*ARC Centre in Bioinformatics, Institute for Molecular Bioscience, The University of Queensland, St. Lucia, 4072, Australia*
[2]*ARC Special Research Centre for Functional and Applied Genomics, Institute for Molecular Bioscience, The University of Queensland, St. Lucia, 4072, Australia*

Received 7 August 2003; Accepted 24 October 2003

Abstract: A new method has been developed for prediction of transmembrane helices using support vector machines. Different coding schemes of protein sequences were explored, and their performances were assessed by crossvalidation tests. The best performance method can predict the transmembrane helices with sensitivity of 93.4% and precision of 92.0%. For each predicted transmembrane segment, a score is given to show the strength of transmembrane signal and the prediction reliability. In particular, this method can distinguish transmembrane proteins from soluble proteins with an accuracy of ~99%. This method can be used to complement current transmembrane helix prediction methods and can be used for consensus analysis of entire proteomes. The predictor is located at http://genet.imb.uq.edu.au/predictors/SVMtm.

 J Comput Chem 25: 632–636, 2004

Key words: SVMtm; transmembrane helix prediction; location of transmembrane segments; coding scheme

Introduction

Transmembrane proteins (TM proteins) represent about 15–30% of the protein sequences in higher eukaryotes, and play important roles across a range of cellular functions.[1] Due to the difficulty in solving their 3D structures by X-ray or NMR, theoretical prediction is important for revealing structures and functions of TM proteins. The goals of transmembrane prediction include (1) differentiation of TM proteins from other proteins, i.e., the soluble proteins, (2) prediction of the locations of TM segments (including accurate prediction of TM segment boundaries), and (3) prediction of the orientation of a TM segment within the membrane. The secondary structure of a membrane-spanning segment can be an α-helix or β-strand, but the TM β-strand usually has fewer residues than an α-helix. Nearly all TM β-strand proteins are found in prokaryotes, and belong to a few protein families. Because of this, we have focused our attention on the prediction of transmembrane α-helices.

Previously, nearly all the computational prediction methods have been based on protein amino acid sequences. A recent comparison of different methods[2] showed that nearly all aspects of prediction need improvement, for example, only around 50% of membrane proteins can be predicted with all segments correct for most methods. Furthermore, accurate differentiation of N-terminal TM helices and signal peptides is still problematic for existing predictors. These shortcomings have a serious impact on the reliability of computational annotation for proteomes. Karin et al.[3] estimated that only 53–59% of all predicted topologies for the proteomes of *Escheria coli, Saccharomyces cerevisiae,* and *Caenorhabditis elegans* were correct. The comparison performed by Chen et al.[2] also showed that no method performed consistently best based on various measures of accuracy. Therefore, the development of new independent methodologies that are able to predict transmembrane segments with advantages on some aspects can complement the current methods and will further strengthen the ability to computationally predict transmembrane segments.

In this work, we develop a new method based on the support vector machine (SVM) approach[4] to predict transmembrane helices. Many membrane prediction methods are based on amino acid hydrophobicity, while advanced methods use different sequence coding schemes. For our method, we compare the performances of coding schemes when using the same SVM model. Through this comparison, we select the best performing coding scheme and implement the algorithm as a predictor. To reflect the reliability of

Correspondence to: Z. Yuan; e-mail: z.yuan@imb.uq.edu.au

Contract/grant sponsor: Australian Research Council.

Contract/grant sponsor: National Institute of Health; contract/grant number: NIDDK DK063400.

a predicted TM segment, a TM score is given as well. A detailed analysis of the relation between score and reliability is performed, and finally, we examine this method's capability of distinguishing membrane proteins from soluble proteins and signal peptides.

Methods

Like neural network (NN), SVM is a well-developed machine-learning algorithm given by Vapnik,[4] with many successful applications in a variety of research areas. The general concept of SVM can be introduced as following. For a two-class problem, there are a series of samples described by the feature vectors x_i ($i = 1, 2, \ldots, N$) with corresponding labels $y_i \in \{+1, -1\}$ ($i = 1, 2, \ldots, N$). In this particular study, the two classes are defined as transmembrane residues (labeled as $+1$) and nontransmembrane residues (labeled as -1). To classify the two classes of samples, SVM learns the boundary regions between samples belonging to two classes by mapping the input samples into a high-dimensional space, and then seeking a separating hyperplane. The hyperplane (determined by coefficient α_i) can be obtained by solving the following optimization problem: Maximize

$$\sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \tag{1}$$

subject to

$$\sum_{i=1}^{N} \alpha_i y_i = 0 \quad (i = 0, 1, 2, \ldots, N \text{ and } 0 \le \alpha_i \le C) \tag{2}$$

where C is a parameter that controls the trade-off between margin and classification error. $K(x_i, x_j)$ is a kernel function, which is used to map the input vectors to a more complicated feature space. In this study, radial basis function (RBF) is selected and given as follows,

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2) \tag{3}$$

The distance of an unlabeled test sample to the hyperplane can be calculated as

$$f(x) = \sum_{i=1}^{N} y_i \alpha_i K(x_i, x_j) + b \tag{4}$$

where b is a constant used to balance the support vector machine outputs. The distance $f(x)$ is used to reflect the propensity of a residue being a transmembrane residue or not. The larger the value, the more likely the residue is a TM residue. It is worth noting that the optimization problem in SVM is a convex optimization problem, which ensures a global optimum. There is no risk of getting stuck to a local minimum, which may occur in gradient-based training of neural networks. To use the prediction function [eq. (4)], we can obtain all the coefficients by using the SVM_Light[5] package and setting $\gamma = 0.1$ [eq. (3)] and $C = 0.5$ [eq. (2)].

The feature vector representing a residue is extracted by the sliding window technique. In a protein sequence, whether a residue belongs to a TM segment or not is determined by its neighboring residues. With a window centered at the residue, 13–20 residues are usually considered. The feature vector for the windowed sequence to represent the residue can be coded in two different ways. First, we adopted the scheming method previously used by the neural network method,[6] in which a residue is coded in a 21-dimensional vector. Within the vector, the first 20 units stand for 20 types of amino acids and the last one represents the break or uncommon amino acid. For convenience, this coding scheme is called the "21-UNIT" method. Because many TM prediction methods are based on hydrophobicity scales, to compare with those methods, protein sequences are also coded by using amino acid hydropathy scales. In a given window, all the residues are replaced by their normalized hydropathy values. When the break or uncommon amino acid occurs, the value is regarded as zero. Three hydropathy scales (JTT, EB, KD) have been used to generate protein profiles. All the scales are normalized with mean zero and standard deviation one.[7] Under this condition, the coding method is simply nominated as the name of the hydropathy scale. One problem to be mentioned is the window size. A large window contains more local sequence information but takes longer for a method to train and test. When selecting the first coding scheme, we found that the window size had very minor impact on the final results, as observed previously when we applied SVMs for protein solvent accessibility prediction.[8] Therefore, the window size is selected as 15 amino acids for the 21-UNIT method. For hydropathy methods, a larger window size is selected and set at 19 amino acids.

Generated by eq. (4), each residue has a real value (transmembrane profile value) and a protein sequence is represented by a series of real values. Based on the transmembrane profile, some algorithms can be used to define TM segments. Dynamic programming has been successfully applied in finding TM segments.[9] MaxSubSeq was an algorithm based on dynamic programming and showed a very promising application in this problem.[10] MaxSubSeq locates the TM segments by maximizing a global score defined as

$$S = \sum_{i=1}^{n} s_i \tag{5}$$

where s_i is the local score for the ith TM segment. n is the number of predicted TM segments, and s_i is defined as,

$$s_i = \sum_{j=k}^{k+m-1} p(i, j) \tag{6}$$

where $p(i, j)$ is the transmembrane profile value at position j in the ith TM segment. m is the length of the TM segment, and is limited to a range of $[l_{\min}, l_{\max}]$. In this work, we set $l_{\min} = 15$ and $l_{\max} = 35$ because nearly all the transmembrane helices have lengths

within this range. n and m can be extracted from a score matrix generated by a recursive algorithm.[10] For the predicted ith TM segment, the s_i is given to show the strength of TM signal.

To examine the performance of our methods, a variety of accuracies were defined. Based on TM segments, specificity (Q_{sp}) is the percentage of correctly predicted segments over the predicted segments and sensitivity (Q_{se}) is the percentage of correctly predicted segments over the true segments. A correctly predicted segment is defined as one that has at least nine residues overlapping the true segment. If all the TM segments in a protein are correctly predicted and the number of predicted segments is correct, it is a correctly predicted protein. Therefore, based on proteins, Q_0 is the percentage of correctly predicted proteins over total proteins. Q_1 is the accuracy when we tolerate one segment prediction error. That means if a protein has one mis-predicted TM segment it can still be regarded as a correctly predicted protein. This could be due to overprediction or underprediction of one TM segment, or correct prediction of the number of segments but misplacement of one segment. Thus, Q_1 can be defined as the percentage of correctly predicted proteins plus proteins with one segment mis-predicted over total proteins. The difference of Q_1 and Q_0 indicates the miss-one-segment errors and the success rate when we overcome this problem.

A nonredundant dataset[11] consisting of 148 well-annotated transmembrane proteins was used to examine our methods. To avoid overestimating the accuracies, the sevenfold crossvalidation test is performed. That means that 148 proteins are divided into seven groups with roughly equal numbers of proteins. Each group is tested after training SVMs on the remaining samples.

Results and Discussions

Table 1 shows the performances of our methods when various coding schemes are used. On all prediction aspects, 21-UNIT is slightly better than the methods coding with hydrophobicity scales. The results indicate that the amino acid hydropathy scales derived from experiments or transmembrane protein datasets may oversimplify the information contained in transmembrane segments, and therefore, give less accurate prediction results. However, the difference is not very significant. Previous observations show that hydrophobicity-based methods are less accurate than advanced methods.[2] We consider the lower accuracy may attribute to two parts, the hydrophobicity scales and the methods used for generation of transmembrane profiles. This can be verified by the following comparison. When the KD scale[12] was selected and the TM profile was a simple average of hydropathy values for each amino acid in a sliding widow, Q_0 could only reach 32%. After MaxSubSeq filtering, Q_0 increased to 51% given by Fariselli et al.[10] In our method, when the support vector machine based on the KD scale is used to generate TM profiles, a significant improvement can be achieved. Strict crossvalidation tests yield the prediction accuracies Q_{sp} = 91.0%, Q_{se} = 92.9%, Q_0 = 60.1% and Q_1 = 86.5%, only slightly lower than the 21-UNIT coding method. The 21-UNIT coding scheme gives the accuracies: Q_{sp} = 92.0%, Q_{se} = 93.4%, Q_0 = 63.5%, and Q_1 = 86.5%. Its performance is better than or comparable to the 28 methods examined by Chen et al.[2] based on a very similar dataset[11] consisting of 165 low-resolution proteins, even if some methods may overpredict their accuracies due to including the testing proteins in their training procedures.

Table 1. Prediction of Transmembrane Segments Based on Different Sequence Coding Schemes.

	Prediction accuracy (%)			
Coding scheme	Q_{sp}	Q_{se}	Q_0	Q_1
21-UNIT	92.0	93.4	63.5	86.5
JTT	91.6	93.0	61.5	83.1
KD	91.0	92.9	60.1	86.5
EB	90.1	92.7	56.1	83.8

Specificity (Q_{sp}) is the percentage of correctly predicted segments over the predicted segments, and sensitivity (Q_{se}) is the percentage of correctly predicted segments over the true segments. Q_0 is the percentage of correctly predicted proteins over total proteins, and Q_1 the percentage of correctly predicted proteins plus proteins with one segment mis-predicted over total proteins.

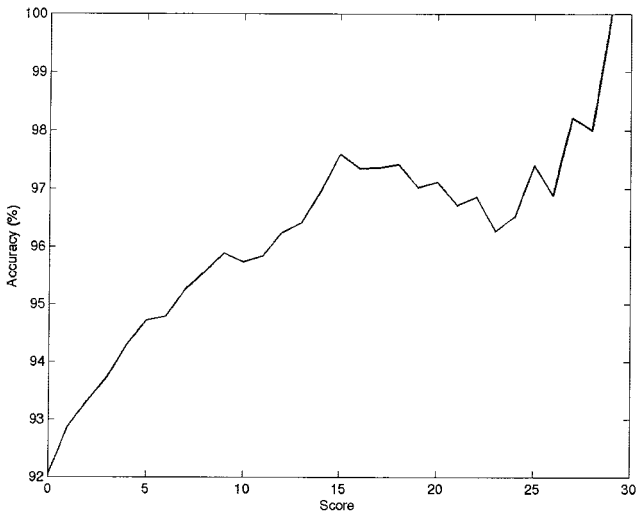

Figure 1. Prediction accuracy of transmembrane segments according to different score thresholds. Transmembrane segments with larger scores are also predicted with more reliability.

To carefully evaluate the reliability of predicted TM segments, we selected different score thresholds and calculated the prediction accuracy for all TM segments with scores larger than the threshold. The results are shown in Figure 1. As the score increases, the accuracy also increases. Fluctuations of accuracy can be observed; however, it can be concluded that prediction accuracies are roughly proportional to the scores. To determine the prediction reliability for a certain score, TM scores are divided into seven groups of ranges; each of them has the same number of TM segments. After calculating the prediction accuracy for each group, the results are given in Table 2. TM segments located in the range (0, 5.6] only have an accuracy of 75.0%. This range covers one-seventh of all the predicted TM segments. A significant in-

Table 2. The Prediction Accuracy for Different Groups of Scores.

Score range	Accuracy (%)
(0, 5.6]	75.0
(5.6, 9.30]	90.4
(9.30, 12.28]	93.3
(12.28, 15.71]	93.3
(15.71, 18.73]	98.0
(18.73, 23.08]	98.1
[23.08, $+\infty$)	96.2

crease in accuracy can be observed for those TM segments with scores higher than 5.6. Six-sevenths of TM segments have scores higher than 5.6 and with accuracy more than 90.0%. Particularly, if the score is larger than 15, the accuracy reaches 96–98% and greater reliability can be achieved. The above analyses indicate that low score TM segments need to be further verified by other prediction methods or biochemical experiments. We illuminate this phenomenon by giving an example of prediction results for protein CYOE_ECOLI (SWISS-PROT ID), shown in Figure 2. It is predicted to have eight TM segments. One segment, starting from residue 160 to residue 174, has a score of 1.9, while the other seven segments all have scores higher than 12. The low score segment actually represents a false positive.

To further discuss this problem, we compared our method with other methods based on lower score TM segments and higher score TM segments. We collected all annotated TM segments (correctly predicted by SVMtm with a score no more than 5.6) and tested them using four transmembrane helix prediction methods: TMHMM2,[13] HMMTOP2,[14] SOSUI,[15] and TopPred2;[16] 80.8, 83.3, 74.4, and 89.7% of the low score TM segments can be correctly predicted by TMHMM2, HMMTOP2, SOSUI, and TopPred2, respectively. When we use higher score TM segments (score larger than 15.7), the accuracies increase to 98.7, 94.4, 97.4, and 95.7%, respectively. The results suggest that the low score TM segments are predicted in lower consensus by different methods. SVMtm can correctly predict some low score TM segments mispredicted by other methods, and therefore, can complement other methods. When we compare different methods based on the 148 membrane proteins, it is worth mentioning that a proportion of them have been used by other authors for developing their methods. Therefore, the comparison cannot be considered strict. Although the accuracies of our method are yielded by strict crossvalidation tests, they are better than or comparable with results of others. The prediction accuracies (Q_{sp}; Q_{se}) for TMHMM2, HMMTOP2, SOSUI, and TopPred2 are (94.3; 89.7%), (90.9; 89.0%), (92.4; 88.7%) and (85.4; 91.2%), respectively. The accuracies for SVMtm are (92.0; 93.4%).

The lower reliability of low score TM segments is a reason for the large difference between Q_0 and Q_1 ($\sim$26% for nearly all coding schemes). The low score segments also include the mispredicted hydrophobic regions of signal peptides as we will discuss later. It is clear that low score TM segments need further attention if we want to improve the prediction performance. The consensus approaches recently developed for TM predictions can reduce the uncertainty of weakly predicted TM segments[17,18] and thus improve the overall accuracy. Using some methods that can accurately distinguish signal peptides and N-terminal transmembrane helices can also reduce the error between Q_0 and Q_1.

Minimizing the false positive and false negative rates is an important task for transmembrane prediction methods. Methods are effective in locating transmembrane segments in real proteins, but they tend to incorrectly identify other hydrophobic clusters in soluble proteins as helical transmembrane segments.[19] The ability to distinguish membrane proteins from soluble proteins is also important for the prediction of protein subcellular localizations. In prediction of subcellular localization, new proteins are initially classified as membrane or nonmembrane proteins because these basic classes have different functions and demand different experimental techniques. Further classification is then performed to predict their destinations in the cell. Here, we select two datasets to examine our methods. One is a nonredundant set of 1993 soluble proteins with known structures and pair-wise identity less than 25%.[20] The other one is the dataset of 1523 signal peptides derived from SignalP predictors.[21] Using the 21-UNIT coding scheme, we ran our prediction methods against three datasets (transmembrane, soluble, and signal peptide) and gathered all the scores for predicted TM segments. In Figure 3A, the score distributions of TM segments from different datasets are compared. It can be found that the predicted TM segments in soluble proteins have lower average scores compared with signal peptides and transmembrane proteins. The TM proteins have the largest average TM scores. These results indicate that the scores provide useful insight into the problem of differentiation among the putative TM segments from transmem-

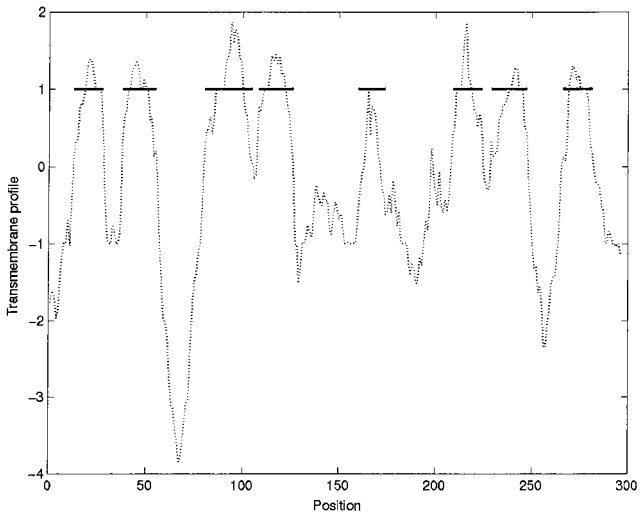

Figure 2. Transmembrane profile and predicted transmembrane segments for protoheme IX farnesyltransferase (SWISS-PROT ID: CYOE_ECOLI). The transmembrane profile is represented by the dashed line while predicted transmembrane segments are represented by solid bars. For each predicted transmembrane segment, its start position, end position, and TM score are listed as (start_end;score). The predicted transmembrane segments are (13_28;13.91), (38_55;15.54), (80_105; 24.40), (108_126;18.70), (160_174;1.9), (209_224;14.91), (229_247; 14.14) and (266_281;12.77). The segment from residue 160 to residue 174 is a false positive with a score of 1.9.

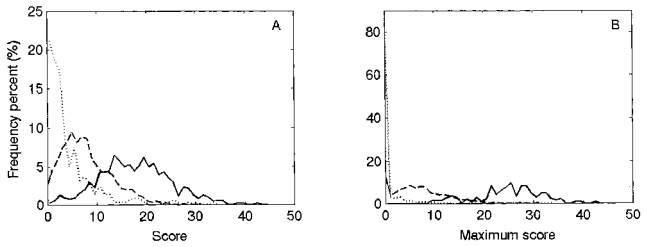

Figure 3. (A) Score distributions of predicted transmembrane segments from soluble proteins (dotted line), signal peptides (dashed line), and transmembrane proteins (solid line). (B) Distribution of maximum scores from soluble proteins (dotted line), signal peptides (dashed line), and transmembrane proteins (solid line). Each protein is represented by its maximum transmembrane score. If no transmembrane segment is predicted, the maximum score is assigned as zero. The comparison in (A) is based on per segment, while the comparison in (B) is based on per protein.

brane proteins, soluble proteins, and signal peptides. To distinguish a TM protein from a soluble protein, we select only the maximum TM score to represent each protein. If a protein has no predicted TM segment, its maximum TM score is set as 0. Figure 3B gives the distributions of maximum TM scores for different datasets. It is obvious that the difference between TM proteins and soluble proteins is enlarged when we only use the maximum TM scores. If the score threshold is set as 10, 98.8% of soluble proteins have maximum scores lower than the threshold, while 98.6% of TM proteins have maximum scores higher than the threshold. If this threshold is also applied to signal peptides, we found 75.2% of these sequences have maximum scores lower than the threshold. Although the information is not accurate enough for differentiation of TM segments and signal peptides, it is obviously a useful feature when developing a differentiation method for N-terminal signal peptides and TM helices.

In conclusion, we have developed a new transmembrane helix prediction method. First, support vector machines and the 21-UNIT coding input scheme are used to generate protein transmembrane profiles. Then, MaxSubSeq is used to define the transmembrane segments. Finally, we filter the predicted TM segments to minimize the false positives from globular proteins and signal peptides. If the maximum TM score is not larger than 10, the protein is reassigned as a soluble protein. For a secreted soluble protein, the signal peptide may be recognized by this filtering step. Specified differentiation methods for N-terminal signal peptides and transmembrane helices are needed to solve this problem.[22,23] The above method has been implemented as a web predictor hosted at http://genet.imb.uq.eud.au/predictors/SVMtm, which can predict single sequence or multiple sequences and SVMtm can be used in combination with other methods for consensus prediction.

Acknowledgments

The authors thank Fasheng Zhang and Adrian Miranda for technical support of the web application and Melissa Davis for helpful discussion.

References

1. Kanapin, A.; Batalov, S.; Davis, M. J.; Gough, J.; Grimmond, S.; Kawaji, H.; Magrane, M.; Matsuda, H.; Schonbach, C.; Teasdale, R. D.; Yuan, Z. Genome Res 2003, 13(6B), 1335.
2. Chen, C. P.; Kernytsky, A.; Rost, B. Protein Sci 2002, 11, 2774.
3. Melen, K.; Krogh, A.; von Heijne, G. J Mol Biol 2003, 327, 735.
4. Vapnik, V. N. The Nature of Statistical Learning Theory; Springer-Verlag: New York, 1995.
5. Joachims, T. In Advances in Kernel Methods—Support Vector Learning; Schölkopf, B.; Burges, C.; Smola, A., Eds.; MIT Press: Cambridge, MA, 1999.
6. Rost, B.; Casadio, R.; Fariselli, P.; Sander, C. Protein Sci 1995, 4, 521.
7. Yuan, Z.; Teasdale, R. D. Bioinformatics 2002, 18, 1109.
8. Yuan, Z.; Burrage, K.; Mattick, J. S. Proteins 2002, 48, 566.
9. Jones, D. T.; Taylor, W. R.; Thornton, J. M. Biochemistry 1994, 33, 3038.
10. Fariselli, P.; Finelli, M.; Marchignoli, D.; Martelli, P. L.; Rossi, I.; Casadio, R. Bioinformatics 2003, 19, 500.
11. Moller, S.; Kriventseva, E. V.; Apweiler, R. Bioinformatics 2000, 16, 1159.
12. Kyte, J.; Doolittle, R. F. J Mol Biol 1982, 157, 105.
13. Krogh, A.; Larsson, B.; von Heijne, G.; Sonnhammer, E. L. J Mol Biol 2001, 305, 567.
14. Tusnady, G. E.; Simon, I. Bioinformatics 2001, 17, 849.
15. Hirokawa, T.; Boon–Chieng, S.; Mitaku, S. Bioinformatics 1998, 14, 378.
16. Claros, M.; von Heijne, G. Comput Appl Biosci 1994, 10, 685.
17. Ikeda, M.; Arai, M.; Lao, D. M.; Shimizu, T. In Silico Biol 2002, 2, 19.
18. Martelli, P. L.; Fariselli, P.; Casadio, R. Bioinformatics 2003, 19(Suppl 1), 1205.
19. Cserzo, M.; Eisenhaber, F.; Eisenhaber, B.; Simon, I. Protein Eng 2002, 15, 745.
20. Noguchi, T.; Akiyama, Y. Nucleic Acids Res 2003, 31, 492.
21. Nielsen, H.; Brunak, S.; von Heijne, G. Protein Eng 1999, 12, 3.
22. Lao, D. M.; Shimizu, T. METMBS Int Conf 2001, p. 119.
23. Yuan, Z.; Davis, M. J.; Zhang, F.; Teasdale, R. D., Biochem Biophys Res Commun 2003, 312, 1278.