W164–W167 *Nucleic Acids Research, 2005, Vol. 33, Web Server issue*
doi:10.1093/nar/gki367

TMBETA-NET: discrimination and prediction of membrane spanning β-strands in outer membrane proteins

M. Michael Gromiha*, Shandar Ahmad[1] and Makiko Suwa

Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), AIST Tokyo Waterfront Bio-IT Research Building, 2-42 Aomi, Koto-ku, Tokyo 135-0064, Japan and [1]Department of Biochemical Engineering and Science, Kyushu Institute of Technology, Iizuka 820 8502, Fukuoka, Japan

Received February 6, 2005; Revised and Accepted February 28, 2005

ABSTRACT

We have developed a web-server, TMBETA-NET for discriminating outer membrane proteins and predicting their membrane spanning β-strand segments. The amino acid compositions of globular and outer membrane proteins have been systematically analyzed and a statistical method has been proposed for discriminating outer membrane proteins. The prediction of membrane spanning segments is mainly based on feed forward neural network and refined with β-strand length. Our program takes the amino acid sequence as input and displays the type of the protein along with membrane-spanning β-strand segments as a stretch of highlighted amino acid residues. Further, the probability of residues to be in transmembrane β-strand has been provided with a coloring scheme. We observed that outer membrane proteins were discriminated with an accuracy of 89% and their membrane spanning β-strand segments at an accuracy of 73% just from amino acid sequence information. The prediction server is available at http://psfs.cbrc.jp/tmbeta-net/.

INTRODUCTION

Discriminating outer membrane proteins from other folding types of globular and membrane proteins is an important task both for identifying outer membrane proteins from genomic sequences and for the successful prediction of their secondary and tertiary structures. Recently, few methods have been reported to identify β-barrel membrane proteins and transmembrane β-barrels in proteomes (1–8). These methods are based on structure based sequence alignment (1), hydrophobicity (2), sequence-profile based HMM (3), amino acid composition in the membrane spanning regions of β-barrel membrane proteins (4), Hidden Markov Model (5–7) and machine learning technique (8). All these methods used minimal information for the analysis and the prediction accuracy is rather modest.

Prediction of membrane spanning β-strands in outer membrane proteins is an important problem to identify the functionally important residues and for modeling their three-dimensional structures. For the past two decades, several methods have been proposed to predict the transmembrane β-strand segments in outer membrane proteins and these approaches include, turn elimination method, amphipathic character of β-strands, hydrophobicity profiles, conformational parameters, sequence alignment and neural network (9–13). However, most of these methods are applicable only to specific types of outer membrane proteins.

In this work, we have systematically analyzed the amino acid compositions of globular and outer membrane proteins using a large dataset and devised a statistical method based on the amino acid composition for discriminating OMPs. We have tested our approach with several sets of globular proteins belonging to four different structural classes, transmembrane helical proteins and OMPs. Our predicted results showed an accuracy of 89% for correctly picking up the OMPs from known annotated sequences and the present method is able to exclude up to 80% of globular proteins and α-helical membrane proteins. Further, we have developed an algorithm based on neural networks for identifying the membrane-spanning β-strand segments using amino acid sequence alone. We observed a good agreement between predicted and experimentally observed transmembrane β-strand segments. A web interface has been set up which allows users to input the sequence and get the type of the protein along with membrane spanning segments automatically. The prediction server is available at http://psfs.cbrc.jp/tmbeta-net/.

*To whom correspondence should be addressed. Tel: +81 3 3599 8046; Fax: +81 3 3599 8081; Email: michael-gromiha@aist.go.jp

MATERIALS AND METHODS

Database

We have used several sets of data for discriminating OMPs: (i) 377 well annotated OMPs (14), (ii) 674 globular proteins belonging to the four structural classes, all-α, all-β, α+β and α/β with the sequence identity of <35% (15), (iii) subset of 27 non-redundant OMP sequences with <35% sequence identity, (iv) non-redundant dataset of 19 OMPs available in Protein Data Bank (16), (v) 377 different sets of data for all the OMPs (for each OMP we constructed a dataset, which does not contain any homologous sequences), (vi) dataset of 85 β-barrel porins, 19 aquaporins and 16 α-helical membrane proteins from Transport Classification Database (17) and (vii) 268 well-annotated α-helical transmembrane proteins (14). Further, we used a dataset of 13 outer membrane proteins of known three-dimensional structures for predicting their membrane spanning β-strand segments. The transmembrane β-strand segments have been assigned from the illustrations provided by the crystallographers in their original papers, who solved the structures. The assignment of β-strands and loops in the membrane are also available in PDB_TM database (18).

Computation of amino acid composition and discrimination of OMPs

The amino acid composition for the set of OMPs ($Comp_{OMP}$) has been computed using the number of amino acids of each type and the total number of residues. It is defined as:

$$\mathrm{Comp}(i) = \frac{\sum n_i}{N} \qquad \mathbf{1}$$

where i stands for the 20 amino acid residues. n_i is the number of residues of each type and N is the total number of residues. The summation is through all the residues in all the considered proteins. The same procedure was repeated for the globular proteins for obtaining their amino acid composition ($Comp_{glob}$). For a new protein, X, firstly, we have calculated the amino acid composition using Equation 1. Then we have calculated the total absolute difference of amino acid composition between protein X and the amino acid composition of globular proteins, and that between protein X and OMPs. The protein X is predicted to be a OMP if the deviation is lowest with $Comp_{OMP}$ and vice versa.

Prediction of membrane spanning β-strand segments

We have constructed a neural network based method for predicting the membrane-spanning β-strand segments (19). In this method, a three-layered neural network with one hidden layer has been used for predictions. Input layer reads the input information about a residue and its sequence neighbors from the neural network through a running window. This input information is then fed forward through linear activation function and the final signal received at the single unit of the output layer is transformed via a sigmoidal function to yield a value between 0 and 1, similar to our previous real value prediction of solvent accessibility (20). This real value of the output unit is regarded as the probability of that residue to be in transmembrane β-strand and final assignment of a residue is made according to a specific cutoff probability.

Back-check, validity check and jack-knife tests

For discriminating OMPs, we have used back-check (self consistency test), validity check, jack-knife and cross-validation methods (21). In back-check prediction, we used the dataset of 377 OMP sequences and 674 globular protein sequences for deriving the amino acid composition and for predicting the type of the protein. For the validity check prediction, we have divided a set of N proteins into equally balanced subsets; parameters are developed on M proteins and then tested on the remaining $N-M$ proteins; the procedure is repeated for all subsets of data to obtain the average accuracy. In jack-knife test, we computed the amino acid composition by omitting one protein and used this information for assigning the type of left-out protein. In cross-validation method, we tested each OMP using the amino acid composition computed on a set that does not contain any homologous sequences (21). For predicting the membrane spanning β-strand segments, we have used 12 proteins (training set) to derive the parameters and the prediction was made for the left-out protein.

RESULTS AND DISCUSSIONS

Discrimination of globular and OMPs

We have calculated the amino acid compositions for 674 globular and 377 OMPs using Equation 1 and the results are presented in Table 1. We observed that the residues Glu, His, Ile, Cys, Gln, Asn, and Ser show subtle difference between the composition of globular and OMPs. While the composition of Glu, His, Ile and Cys are higher in globular proteins than OMPs an opposite trend is observed for Ser, Asn and Gln. Detailed analysis on the three-dimensional structures of OMPs showed that the residues Ser, Asn and Gln play an important role to the formation of β-barrel structures in the membrane, stability and function of OMPs (21).

Table 1. Amino acid composition for the 20 amino acid residues in globular and OMPs

Residue	Composition (%)	
	Globular	OMP
Ala	8.47	8.95
Asp	5.97	5.91
Cys	**1.39**	**0.47**
Glu	**6.32**	**4.78**
Phe	3.91	3.68
Gly	7.82	8.54
His	**2.26**	**1.25**
Ile	**5.71**	**4.77**
Lys	5.76	4.93
Leu	8.48	8.78
Met	2.21	1.56
Asn	**4.54**	**5.74**
Pro	4.63	3.74
Gln	**3.82**	**4.75**
Arg	4.93	5.24
Ser	**5.94**	**8.05**
Thr	5.79	6.54
Val	7.02	6.76
Trp	1.44	1.24
Tyr	3.58	4.13

The amino acid residues that have high (>0.9) difference between OMP and globular proteins are highlighted in bold.

For a new protein, we have calculated the amino acid composition and the total deviation in the amino acid compositions of 20 amino acid residues from the average values (globular proteins and OMPs) given in Table 1. As an example, for adenovirus DNA binding protein (1ADT), the deviation of amino acid composition from globular protein is 34.18, which is less than that of OMP (39.89). Hence this protein is predicted as a non-OMP. On the other hand, for OmpA protein, the deviation from OMP (36.66) is less than that from globular protein (37.78) and hence it is identified as an OMP (Figure 1). These results show the excellent agreement with experimental observations.

We have correctly identified 334 out of 377 OMPs (89%) and excluded 531 of 674 globular proteins (79%). The validity-check method described earlier yielded an average accuracy of 84 and 78%, respectively, for correctly identifying OMPs and excluding globular proteins. The cross-validation method tested for each OMP using the frequency computed with a set of non-homologous sequences showed an accuracy of 80% for correctly assigning OMPs. We have carried out a jack-knife test using a dataset of 27 non-redundant OMP sequences and the OMPs were discriminated at an accuracy of 93%. Further, we have successfully eliminated 213 of the 268 α-helical transmembrane proteins (80%) and correctly identified 18 of the 19 OMPs of known structure (95%). Moreover, the present method performed very well in discriminating different sets of β-barrel porins, α-helical membrane proteins and aquaporins from Transport Classification Database (17). It correctly identified 91% of the 85 β-barrel porins and excluded 74% of 19 aquaporins and 88% of 16 α-helical membrane proteins. These results show that the method based on amino acid composition could discriminate OMPs with accuracy in the range 80–95%, which is comparable to other methods in the literature (21).

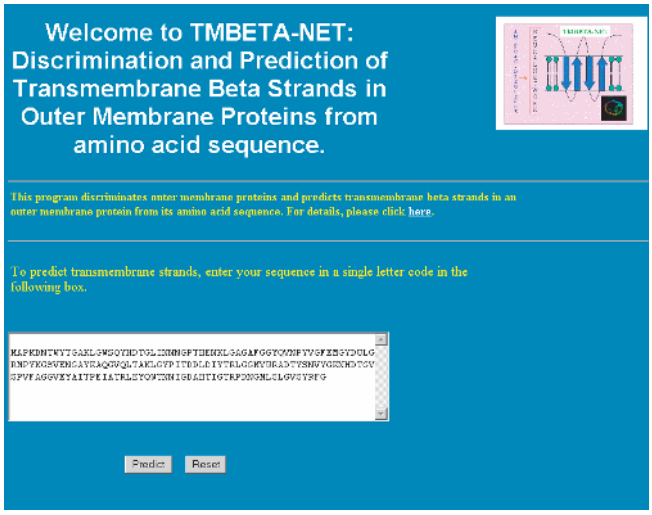
Prediction of membrane spanning β-strands

We have trained a set of 12 proteins and predicted the membrane spanning β-strands in the left-out protein. Our method predicts with an average accuracy of 73.2% for the 13 considered proteins, without any additional information. Further, our algorithm has been tested with 20 non-redundant β-barrel membrane proteins listed in Bagos *et al*. (22) and we obtained similar accuracy. The average correlation of 0.4644 is comparable to or better than the values obtained in several structure prediction methods and/or two-state solvent accessibility prediction algorithms used for globular proteins (23,24). Further, the performance of our network is highly convincing due to the narrow range (0.3 to 0.6) of correlation obtained for each of the protein. We have also computed the sensitivity and specificity for each protein and the values lies in the range 0.5–0.8 for sensitivity and 0.7–0.9 for specificity. These results emphasize the reliability of our method with high confidence. The work on alignment profiles is on progress.

Prediction on the web

We have developed a web server for discriminating outer membrane proteins and predicting the membrane spanning β-strands. Figure 1a shows the details of our web server including the input options. It takes the amino acid sequence in one letter format as the input and automatically omits gaps

(a)



(b)

Results of composition based discrimination of Outer Membrane Proteins:

Amino acid composition, deviation from Globular proteins (GD)
and deviation from Outer Membrane Proteins (OD) in the sequence are as follows:

Residue	Composition (%)	GD (%)	OD (%)
A	6.98	1.49	1.97
C	0.00	1.39	0.47
D	5.81	0.16	0.10
E	3.49	2.83	1.29
F	2.33	1.58	1.35
G	15.12	7.30	6.58
H	2.33	0.07	1.08
I	4.07	1.64	0.70
K	4.07	1.69	0.86
L	6.40	2.08	2.38
M	2.91	0.70	1.35
N	6.98	2.44	1.24
P	4.65	0.02	0.91
Q	2.91	0.91	1.84
R	3.49	1.44	1.75
S	3.49	2.45	4.56
T	8.14	2.35	1.60
V	5.81	1.21	0.95
W	2.91	1.47	1.67
Y	8.14	4.56	4.01
Total	-	37.78	36.66

Amino acid composition of this sequence seems to be similar to outer-membrane proteins.

(c)

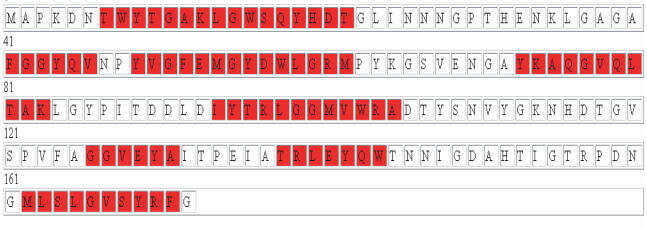


Figure 1. Web based prediction of membrane-spanning β-strands, (**a**) First page showing the input format (amino acid sequence in single letter code; an example is shown for 1bxw, outer membrane protein A), (**b**) Amino acid composition, deviation from globular and OMP, and the type of the protein. OmpA is correctly predicted as an outer membrane protein. (**c**) The stretch of predicted amino acid residues in membrane spanning β-strands, highlighted in red.

and numbers. The output formats are shown in Figure 1b and c. In Figure 1b, we show the amino acid composition of the protein and the deviation from the compositions of globular and membrane proteins. The type of the protein based on amino acid composition is also indicated. As an example, for OmpA protein, the deviations from globular and OMPs are, respectively, 37.37 and 36.68, and the protein is identified as an OMP. Figure 1c displays the transmembrane β-strand segments as a stretch of amino acid residues highlighted in red and OmpA shows the presence of eight membrane spanning β-strand segments. Further, the information about the probability of each residue to be in membrane spanning β-strand have been provided with a color-coding scheme, which may be useful for the users to identify the membrane-spanning β-strands with various confidence levels. The prediction results are freely available at http://psfs.cbrc.jp/tmbeta-net/.

CONCLUSIONS

We have devised a method based on amino acid composition for discriminating outer membrane proteins from other folding types of globular and membrane proteins. A neural network based prediction algorithm has been developed for identifying the membrane spanning β-strands in outer membrane proteins. The reliability of the present method has been examined with several sets of data and different measures such as, accuracy, correlation, sensitivity and specificity. We observed a good agreement between predicted results and experimental observations both in discrimination and prediction. A web server has been developed for the prediction purpose and the results are available online, which may be very helpful for the users to discriminate outer membrane proteins as well as to get the membrane spanning β-strands for an outer membrane protein.

ACKNOWLEDGEMENTS

We thank Dr. Tomotsu Noguchi for providing the list of outer membrane protein structures and Dr. Yutaka Akiyama for encouragement. Funding to pay the Open Access publication charges for this article was provided by the Ministry of Education, Sports, Science and Technology of Japan.

Conflict of interest statement. None declared.

REFERENCES

1. Gnanasekaran,T.V., Peri,S., Arockiasamy,A. and Krishnaswamy,S. (2000) Profiles from structure based sequence alignment of porins can identify beta stranded integral membrane proteins. *Bioinformatics*, **16**, 839–842.
2. Wimley,W.C. (2002) Toward genomic identification of beta-barrel membrane proteins: composition and architecture of known structures. *Protein Sci*., **11**, 301–312.
3. Martelli,P.L., Fariselli,P., Krogh,A. and Casadio,R. (2002) A sequence-profile-based HMM for predicting and discriminating beta barrel membrane proteins. *Bioinformatics*, **18**, S46–S53.
4. Liu,Q., Zhu,Y., Wang,B. and Li,Y. (2003) Identification of beta-barrel membrane proteins based on amino acid composition properties and predicted secondary structure. *Comput. Biol. Chem*., **27**, 355–361.
5. Bigelow,H.R., Petrey,D.S., Liu,J., Przybylski,D. and Rost,B. (2004) Predicting transmembrane beta-barrels in proteomes. *Nucleic Acids Res*., **32**, 2566–2577.
6. Bagos,P.G., Liakopoulos,T.D., Spyropoulos,I.C. and Hamodrakas,S.J. (2004) A Hidden Markov Model method, capable of predicting and discriminating beta-barrel outer membrane proteins. *BMC Bioinformatics*, **5**, 29.
7. Bagos,P.G., Liakopoulos,T.D., Spyropoulos,I.C. and Hamodrakas,S.J. (2004) PRED-TMBB: a web server for predicting the topology of beta-barrel outer membrane proteins. *Nucleic Acids Res*., **32**, W400–W404.
8. Natt,N.K., Kaur,H. and Raghava,G.P. (2004) Prediction of transmembrane regions of beta-barrel proteins using ANN- and SVM-based methods. *Proteins*, **56**, 11–18.
9. Paul,C. and Rosenbusch,J.P. (1985) Folding patterns of porin and bacteriorhodopsin. *EMBO J*., **4**, 1593–1597.
10. Vogel,H. and Jahnig,F. (1986) Models for the structure of outer-membrane proteins of Escherichia coli derived from raman spectroscopy and prediction methods. *J. Mol. Biol*., **190**, 191–199.
11. Welte,W., Weiss,M.S., Nestel,U., Weckesser,J., Schiltz,E. and Schulz,G.E. (1991) Prediction of the general structure of OmpF and PhoE from the sequence and structure of porin from Rhodobacter capsulatus. Orientation of porin in the membrane. *Biochim. Biophys. Acta*, **1080**, 271–274.
12. Gromiha,M.M. and Ponnuswamy,P.K. (1993) Prediction of transmembrane beta-strands from hydrophobic characteristics of proteins. *Int. J. Pept. Protein. Res*., **42**, 420–431.
13. Gromiha,M.M., Majumdar,R. and Ponnuswamy,P.K. (1997) Identification of membrane spanning beta strands in bacterial porins. *Protein Eng*., **10**, 497–500.
14. Gardy,J.L., Spencer,C., Wang,K., Ester,M., Tusnady,G.E., Simon,I., Hua,S., deFays,K., Lambert,C., Nakai,K. *et al*. (2003) PSORT-B: Improving protein subcellular localization prediction for Gram-negative bacteria. *Nucleic Acids Res*., **31**, 3613–3617.
15. Wang,X. and Yuan,Z. (2000) How good is prediction of protein structural class by the component-coupled method? *Proteins*, **38**, 165–175.
16. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res*., **28**, 235–242.
17. Busch,W. and Saier,M.H.,Jr (2002) The transporter classification (TC) system, 2002. *Crit. Rev. Biochem. Mol. Biol*., **37**, 287–337.
18. Tusnady,G.E., Dosztanyi,Z. and Simon,I. (2005) PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank. *Nucleic Acids Res*., **33**, D275–D278.
19. Gromiha,M.M., Ahmad,S. and Suwa,M. (2004) Neural Network Based Prediction of Transmembrane β-strand Segments in outer membrane proteins. *J. Comp. Chem*., **25**, 762–767.
20. Ahmad,S., Gromiha,M.M. and Sarai,A. (2003) Real value prediction of solvent accessibility from amino acid sequence. *Proteins*, **50**, 629–635.
21. Gromiha,M.M. and Suwa,M. (2005) A simple statistical method for discriminating outer membrane proteins with better accuracy. *Bioinformatics*, **21**, 961–968.
22. Bagos,P.G., Liakopoulos,T.D. and Hamodrakas,S.J. (2005) Evaluation of methods for predicting the topology of beta-barrel outer membrane proteins and a consensus prediction method. *BMC Bioinformatics*, **6**, 7.
23. Rost,B. and Sander,C. (1994) Conservation and prediction of solvent accessibility in protein families. *Proteins*, **20**, 216–226.
24. Naderi-Manesh,H., Sadeghi,M., Arab,S. and Moosavi Movahedi,A.A. (2001) Prediction of protein surface accessibility with information theory. *Proteins*, **42**, 452–459.