Online Discrimination of β-Barrel Membrane Proteins from Amino Acid Sequence

M. Michael Gromiha michael-gromiha@aist.go.jp

Makiko Suwa m-suwa@aist.go.jp

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

Keywords: outer membrane protein β-strand, amino acid composition, statistical analysis, neural network, support vector machine

1 Introduction

Outer membrane proteins (OMPs or β-barrel membrane proteins) perform a variety of functions, such as mediating non-specific, passive transport of ions and small molecules, selectively passing the molecules like maltose and sucrose and are involved in voltage dependent anion channels. These proteins contain β-strands as their membrane spanning segments and are found in the outer membranes of bacteria, mitochondria and chloroplast. The assembly of OMPs is somewhat more complex when compared to the assembly of transmembrane helical proteins having α-helices as transmembrane parts. This is probably due to the difference of amino acid sequences in the transmembrane part strands and helices; transmembrane helical proteins contain a stretch of hydrophobic amino acid residues whereas transmembrane strand proteins are intervened by several charged and polar residues. Because of this feature, most predictive schemes, which are successful in predicting transmembrane helical segments, fail to predict the transmembrane strand segments and discriminating β-barrel membrane proteins.

Discriminating β-barrel membrane proteins from other folding types of globular and membrane proteins is an important task both for identifying β-barrel membrane proteins from genomic sequences and for the successful prediction of their secondary and tertiary structures. We have devised statistical methods and machine learning techniques based on the composition of amino acid residues and residue pairs for discriminating OMPs. We have tested our approach with several sets of globular proteins belonging to different structural classes, folding types, transmembrane helical proteins, and OMPs obtained from both well annotated sequences and known three dimensional structures. The present method showed an accuracy of 91% for correctly picking up the OMPs from known annotated sequences and 95% for excluding globular and α-helical membrane proteins. These accuracy levels are higher than other methods in the literature.

2 Construction of dataset

We have constructed several sets of data for the discrimination of OMPs: (i) a dataset of 377 well annotated OMPs obtained from PSORT database and a subset of 208 non-redundant OMP sequences with less than 40% sequence identity obtained with CD-HIT algorithm, (ii) non-redundant dataset of 19 known OMP structures with the sequence identity of less than 25%, (iii) 674 globular proteins belonging to different structural classes (155 all-α, 156 all-β, 184 α+β and 179 α/β proteins), (iv) non-redundant data set of 1602 globular proteins belonging to 30 different folds obtained from Protein Data Bank (v) a dataset of 268 well-annotated transmembrane α-helical proteins and a subset of 206 non-redundant transmembrane helical proteins obtained from PSORT and (vi) 85 β-barrel porins, 19 aquaporins and 16 α-helical membrane proteins from Transport Classification Database (TCDB).

3 Discrimination algorithms: Statistical and machine learning methods

The amino acid composition for the set of OMPs and globular proteins has been computed using the expression: $Comp(i) = \Sigma\ n_i/N$, 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. For a new protein, X, firstly, we have calculated the amino acid composition. 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 ($Comp_{OMP}$). The protein X is predicted to be OMP if the deviation is lowest with CompOMP and vice versa.

We have also used amino acid pair preference and motifs for discrimination of OMPs. In these methods, we have calculated the dipeptide (motif) composition of globular and OMP using the equation, $Dipep(i,j) = \Sigma N_{ij} / (\Sigma N_i + \Sigma N_j)$, where i,j stands for the distribution of 20 amino acid residues at positions i and i+1 (i+2 for motifs). $N_{i,j}$ is the number of residues of type i followed by the residue j. ΣN_i and ΣN_j are the total number of residues of type i and j, respectively, the difference between them ($\sigma_{OMP\text{-}glob}$). For a new protein, we computed the composition and given weights to the dipeptide (motif) composition with $\sigma_{OMP\text{-}glob}$; (iii) calculated the sum of weighted dipeptide (motif) composition and (iv) the protein X is predicted to be an OMP if the total weighted dipeptide (motif) composition is positive and globular protein otherwise.

Further we have used support vector machines, neural networks and other machine learning techniques for discriminating OMPs. The performances of these methods have been assessed with cross-validation techniques.

4 Discrimination of OMPs

The statistical method based on amino acid composition shows the accuracy of 84% and 78% respectively, for correctly identifying OMPs and excluding globular proteins [1]. The performance of the methods based on residue pair preference and motifs is better than that of amino acid composition [2]. In SVM, we have combined the compositions of 18 amino acids and 10 residue pairs and obtained the accuracy of 91% for correctly identifying OMPs and 95% for excluding other folding types of globular and membrane proteins [3]. Neural network method based on amino acid composition could distinguish the OMPs and globular proteins at an accuracy of 91% and it could also discriminate 1602 globular proteins from 30 different folding types to an accuracy of 95%.

5 Discrimination on the web

We have developed web servers for discriminating β-barrel membrane proteins from amino acid sequence [4]. The users have the feasibility of selecting the method for discrimination. In the output we provide the amino acid composition/ residue pair composition/ motif composition of the query sequence and the discrimination result based on the composition of the selected method. The discrimination results are available at http://psfs.cbrc.jp/tmbetadisc/.

References

[1] Gromiha, M. M. and Suwa, M., A simple statistical method for discriminating outer membrane proteins with better accuracy, *Bioinformatics,* 21:961-968, 2005.

[2] Gromiha, MM., Ahmad, S., and Suwa, M. Application of residue distribution along the sequence for discriminating outer membrane proteins, *Comput. Biol Chem*. 29:135-142, 2005.

[3] Park, K-J., Gromiha, M.M., Horton, P. and Suwa, M., Discrimination of outer membrane proteins using support vector machines, *Bioinformatics* (in press), 2005.

[4] Gromiha, M.M., Ahmad, S., and Suwa, M., TMBETA-NET: Discrimination and prediction of membrane spanning β-strands in outer membrane proteins, *Nucleic Acids Res*., 33, W164-167, 2005.