Discrimination of β-Barrel Membrane Proteins Using Machine Learning Techniques

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, residue pair preference, amino acid properties, neural network, support vector machine

1 Introduction

β-barrel membrane proteins (TMBs) perform a variety of functions in living organisms and these proteins contain β-strands as their membrane spanning segments. The membrane spanning segments of TMBs contain several charged and polar residues in contrast with a stretch of hydrophobic amino acid residues in transmembrane helical (TMH) proteins. Hence, most predictive schemes, which are successful in discriminating TMH proteins, fail to discriminate 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. In recent years, machine learning techniques are widely used in different areas of Bioinformatics, such as, predicting protein secondary structures, solvent accessibility, fold recognition etc. In this work we have analyzed several machine learning techniques for discriminating β-barrel membrane proteins and utilized the methods, neural networks and support vector machines based on the composition of amino acid residues, residue pair preferences and amino acid properties for discriminating TMBs. We obtained an accuracy of 94% for correctly discriminating TMBs and globular/TMH proteins, which is better 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) 674 globular proteins belonging to different structural classes (155 all-α, 156 all-β, 184 α+β and 179 α/β proteins), (iii) non-redundant data set of 1602 globular proteins belonging to 30 different folds obtained from Protein Data Bank, (iv) a dataset of 268 well-annotated transmembrane α-helical proteins and a subset of 206 non-redundant transmembrane helical proteins obtained from PSORT.

3 Features Used in Machine Learning Methods

We have used amino acid composition, residue pair preference and amino acid properties for discriminating TMBs with machine learning techniques. The amino acid composition for the set of TMBs and globular proteins has been computed using the expression [1]: \( \text{Comp}(i) = \sum \frac{n_i}{N} \), where \( n_i \) 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 amino acid pair preference is calculated using the equation [2], \( \text{Dipep}(i,j) = \frac{\Sigma N_{ij}}{\Sigma N_i + \Sigma N_j} \), where \( i,j \) stands for the distribution of 20 amino acid residues at positions \( i \) and \( i+1 \). \( N_{ij} \) is the number of residues of type \( i \) followed by the residue \( j \). \( \Sigma N_i \) and \( \Sigma N_j \) are the total number of residues of type \( i \) and \( j \), respectively.

The total amino acid property for each residue \( j \) in each TMB \( i \) was computed using the standard
formula [3], \( P_{\text{total}}(i,j) = \sum P(i,j) \), where, \( P(i,j) \) is the property value of \( j \)th residue for the \( i \)th TMB and the summation is over \( n \), the total number of residue type \( j \) in a protein. We have carried out the computations for a set of 49 physico-chemical, energetic and conformational properties of amino acid properties [4] and used them for discrimination.

4 Discrimination of OMPs

We have analyzed the performance of several machine learning algorithms, such as Bayseian network, logistic function, neural networks, support vector machines, k-nearest neighbor, bagging, classification via regression and decision tree for discriminating TMBs [5]. We observed that the neural networks discriminated the TMBs with an accuracy of 91% using amino acid composition. The support vector machines improved the accuracy up to 93.9% using the composition of 18 selected amino acid residues and 10 residue pairs [6]. Recently we used the amino acid properties as the attributes and the neural network method raised the accuracy to 94.4% for discriminating TMBs. Further, it could also exclude 1612 globular proteins from 30 different folding types to an accuracy of 99.4% (Figure 1).

![Figure 1: The accuracy of excluding 30 major folding types of globular proteins. The CATH classification is used to denote the folding types.](image)

References


