Int J Pharm Bio Sci 2014 Oct; 5(4): (B) 621 - 634

Research Article **Bioinformatics**

International Journal of Pharma and Bio Sciences **ISSN 0975-6299**

ROLE OF AMINO ACID PROPERTIES TO DISCRIMINATE THE TRANSMEMBRANE PROTEIN STRUCTURE SEGMENTS

T. DHARMA RAO, C. VIGNESH AND K. SARABOJI*

Protein Crystallography Laboratory, Department of Bioinformatics, School of Chemical and Biotechnology, SASTRA University, Thanjavur - 613 401, India.

ABSTRACT

Membrane proteins play important roles in the biological process and accurate discrimination of membrane proteins from non-membrane proteins/globular ones would help to locate them in the genome sequences. However, the structural biology of membrane protein is limited due to the physiochemical complexities in determining its three dimensional structures. Thus the requirement of predicting membrane protein structure from sequence is increased and become a central problem in molecular biology; interestingly, several computational strategies and discriminating parameters were developed for the successful prediction of membrane protein structures. Studies have been reported that the transmembrane helical proteins could be discriminated with the accuracy of 90%, reflects the predation strength of the present algorithms. However, this accuracy fluctuates with other class of membrane proteins indicates the need for better physico-chemical observations of the specific folds. Thus, here performed a preliminary systematic analysis to study the role of various physico-chemicals, energetic and conformational amino acid properties to discriminate the transmembrane (TM) and non-transmembrane (NTM) segments of α and β class membrane proteins of diverse superfamily. The present study suggests the superfamily based discriminant properties to identify the transmembrane regions. We found that average numbers of surrounding residues, number of long-range contacts and total non-bonded energy can discriminate membrane proteins of all the superfamilies in the dataset with the accuracy of 85-91%. Thus we suggest the addition of these parameters can improve the accuracy of prediction for the specific superfamily.

KEY WORDS: Membrane proteins, Protein Superfamily, Amino acid property, Protein structure prediction.

K. SARABOJI
Protein Crystallography Laboratory, Department of Bioinformatics, School of Chemical and Biotechnology, SASTRA University, Thanjavur - 613 401, India.

INTRODUCTION

Transmembrane proteins are the integral membrane proteins that spans through the entire biological membrane i.e., phospholipid bilayer which forms the hydrophobic core on both the sides of the membrane[1]. Since the membrane proteins play a vital role in several cellular functions and processes such as cell signalling, transportation of ions across the membranes, cell adhesion, immune response, nutrient absorption, intercellular communication etc [2,3], they are treated as the targets of over 50% of drugs in pharmaceutical industry which have the major impact in diseases like cancer, heart diseases, Alzheimer's etc[4,5]. Interestingly α-helical structured transmembrane proteins roughly cover 30% of the entire genome and play a significant role in cellular and biological functions[6] and they are mostly found in the cytoplasmic membranes of all living organisms[7]. Whereas transmembrane β-barrel structured proteins were located in the outer membranes of Gram-negative bacteria, mitochondria and chloroplasts, with a large coiled sheet that forms pore in the membrane. Compared to α-helical membrane proteins, β barrel proteins are less frequent[8] but found as potential drug targets and plays a role in bacterial virulence[7,9]. Despite their biological and pharmacological importance, they are underrepresented in databases[10]; but the detailed knowledge of three dimensional structures and the topology of membrane proteins are critically required to understand their functionality. In recent years, the difficulties in the determination of the three dimensional structure of membrane proteins using crystallographic technique open the doors to develop computational strategies and improve the methods to predict the membrane proteins structures[11,12] which are very important in biological research. Various methodologies have been implemented to predict the transmembrane regions of the membrane protein and the studies have shown that the amino acid distribution patterns are important identifiable features for α-helical proteins: transmembrane helices of long apolar stretches and the distribution of charged residues in the polar regions[13-14,16]. Further, amino acid propensities are demonstrated the important role to indentify, whether the protein has multi-spanning or single-spanning architecture. Particularly the inside/outside helix caps of multi-spanning helices are richer in Phe and Trp residues whereas Cys, Gly, His, Leu, Lys, Phe, Pro,Ser, Thr, Tyr and Val show preference for single spanning helix caps[17]. Of the various strategies, amino acid hydrophobicity is found as potential measure to predict the membrane bound regions[13,18-20]. Further, statistical methods utilize the hydrophobicity, amphiphilicity and polarity information of the amino acids to categorize the transmembrane segments in a better way[21]. As an outcome, several tools such as TopPred[22], SOSUI[23], MEMSAT[25,] etc., were developed using hydrophobicity analysis as base method to predict the transmembrane regions with the maximum accuracy of 83%. However this strategy is useful to identify the long stretches of hydrophobic residues in α-helices, but it is difficult to predict the outer membrane beta proteins due to less hydrophobic motifs and higher length variation[25]. In addition, the approaches based on structural properties such as inter-helical interactions which are commonly found within the transmembrane helices[6], dynamic programming based sequence alignment[26] and contact prediction methods[27] were developed to improve the transmembrane prediction accuracy. Interestingly, several statistical methods utilizing Hidden Markov Model (HMM) improves prediction accuracy[9,14,28,29]; however these methods were sensitive to the sequence length of the protein[5]. A neural network method developed by Martelli et al.[30] using 12 outer membrane proteins (OMP) predict the OMPs with the accuracy of 84%. Interestingly, Gromiha and co-workers devised a method based on neural networks, which can predict the membrane spanning beta-strand segments with an accuracy of 89% for α/β class proteins and 73% for all-β class of proteins[31]. A stretch of hydrophobic amino acids commonly found in signal peptides and transmembrane helices makes the complex situation to several prediction strategies[32]. However, the attempts were performed to combine the integrated signal peptide prediction along with the evolutionary information[33,34] to improve transmembrane

topology prediction; interestingly a significant improvement (accuracy of 89%) was achieved when it includes the support vector machines[35]. In addition to the *ab initio* methods, template based modelling methods, show success in producing higher accuracy structural models of any type or size[36]; but these methods involves the difficulty of handling lipid environment[37]. Gromiha and Suwa[38] analyzed the influence of various amino acid properties for discriminating OMPs using different machine learning algorithms and observed that most of the properties have discriminated the OMPs with the accuracy of 94%. Recently, Nugent and Jones[39] developed a *de novo* method to predict the large membrane protein structures using fragment-assembly and correlated mutation analysis. Overall, the accuracy of strategies of theoretical structure prediction algorithms depends on the structural class of the proteins and fold[40], which indicates the requirement of further optimization of the existing algorithms. With this background, as the physical and chemical properties of the amino acids are the important factors to shape the protein and function, the present study aims to examine the role of various amino acid properties which can discriminate the transmembrane (TM) and non-transmembrane segments (NTM) of α and β class membrane proteins at the superfamily level. This study suggests the significant parameters which can help to improve the existing membrane structure prediction strategies.

MATERIALS AND METHODS

Dataset

A non-redundant dataset of 292 α and 71 β-class membrane proteins of various resolution were taken for the analysis from Protein Data Bank of Transmembrane Proteins (PDBTM)[41]. Further, the dataset is classified based on the superfamily, which yields 122 α-class and 33 β-class proteins which belongs to 10 and 4 different superfamilies respectively; whereas the superfamilies found with only one membrane protein structure are removed from the dataset. The transmembrane and non-transmembrane regions of the proteins in the dataset were identified using TMDET algorithm[42] based on locating the spatial positions of transmembrane proteins in lipid bilayer using their 3D atomic coordinates. The final dataset used in the present study is shown in Table 1.

Calculation of physico-chemical properties of amino acids

We have computed a set of various physico-chemical, energetic and conformational amino acid properties using the normalized values provided in Gromiha et al.[43]. These properties have been successfully demonstrated to understand the folding and stability of proteins. In the present study, we have analyzed 30 different properties after eliminating some of the properties which are specific to globular proteins such as compressibility, solvent accessible reduction ratio, etc. The properties were calculated for the transmembrane and non-transmembrane regions separately for each protein in the training dataset using,

$$P = \sum_{i=1}^{20} \frac{p(i).n(i)}{N} \qquad \dots \quad (1)$$

where *p(i)* and *n(i)* are the property values of the *i* th amino acid and the number of amino acids of *i* th type in a protein respectively. *N* is the total number of residues in a protein.

Table 1
Dataset of α and β class membrane proteins based on their superfamilies. The RCSB Protein Data Bank (PDB) ids are also given.

S.No	Superfamilies and PDB ids	No. of Proteins
	α-class membrane proteins	
1	Rhodopsin-like receptors and pumps *(PDB ids: 1XIO, F93, 2F95, 2K9P, 2LNL, A7K, 3AM6, 3AYM, 3DDL, 3SN6, 3UG9, 3V2W, 4EA3, 4EIY, 4FPD, 4GRV, 4H33, 4HYJ, 4JKV, 4KNF, 4L6R, 4MBS, 4MQS)*	23
2	Ion channel superfamily *(PDB ids: 1LNQ, 1ORQ, 1P7B, 3J5P, 3JYC, 3PJS, 3PJZ, 3UKM, 3VOU, 4BGN, 4F35, 4GX0, 4I9W,4J7C,4JTA,4LP8,4LTO)*	17
3	Photosynthetic reaction centres and photosystems *(PDB ids: 2AXT, 2WSC, 3A0B, 3AOU, 3ARC, 3BZ1, 4AC5, 4FE1, 4IL6, 4IN5, 4IXQ, 4JC9, 4J72)*	13
4	Electron transport chain complex IV *(PDB ids: 1AR1, 1FFT, 1M56, 1QLE, 2Y69, 2YEV, 3ABK, 3AYF, 3BVD, 3DTU, 3EH3, 3MK7, 3O0R)*	13
5	Major intrinsic protein *(PDB ids: 1FX8, 1J4N, 2EVU, 2W1P, 3C02, 3CN5, 3IYZ, 3LLQ, 3Q7K, 3TDO, 4FC4)*	11
6	APC superfamily *(PDB ids: 2A65, 2JLN, 2WSW, 2XQ2, 3GI8, 3L1L, 3QE7, 4C7R, 4DJI, 4M48)*	10
7	Major facilitator superfamily *(PDB ids: 4LEP, 4LDS, 4J05, 4IU8, 4IKV, 3WDO, 3O7P, 2XUT, 2GFP, 1PW4)*	10
8	Electron transport chain complex III *(PDB ids: 1BCC, 1BE3, 1Q90, 2A06, 2D2C, 2FYN, 2IBZ, 2YIU, 3CX5)*	9
9	p-type ATPase *(PDB ids: 1MHS, 3A3Y, 3B8C, 3IXZ, 3J09, 4BBJ, 4H1W, 4HQJ, 4NAB)*	9
10	Receptor type kinases *(PDB ids: 2JWA,2K1K, 2K9Y, 2L2T, 2L6W, 2LZL, 2M20)*	7
	β-class membrane proteins	
11	Ligand gated protein channels *(PDB ids: 1FEP, 1KMO, 1XKW, 2GRX, 2GUF, 2HDF, 2IAH, 3CSL, 3EFM, 3FHH, 3QLB, 3V89, 4AIQ, 4EPA)*	14
12	Trimeric Proteins *(PDB ids: 4GEY, 3WI4, 3HW9, 3A2S, 2POR, 2O4V, 1H6S, 1E54)*	8
13	Autotransporters *(PDB ids: 1UYN, 3KVN, 3QQ2, 3SLJ, 4E1T)*	5
14	oprD/algE superfamily *(PDB ids: 3JTY, 3RBH, 3SY9, 3SYB, 4FRX, 4FSP)*	6

Statistical test for significance

The mean and the standard deviation of each property in the specific class (α or β) was used as reference to compute the significant difference. The significant properties, which can discriminate the transmembrane regions, were identified using the difference observed between property values at TM and NTM regions. Here, we have used T-test to find the significance of difference, at the class level (α or β), using the formula,

$$Z = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \qquad \ldots \quad (2)$$

where, $\bar{X}$ is the sample mean, n stands for number of samples, μ_0 represents distribution mean and S represents standard deviation.

Superfamily based discrimination

The significant properties, obtained based on T-test analysis, were further examined at the superfamily level. The best three properties which show more deviation between transmembrane and non-membrane segments were used for further discrimination; and those properties are considered as 'discriminant' properties.

Validation

The observed discriminant properties are validated at the superfamily level using a test data comprising three proteins of the same superfamily. Further, each protein is divided into the short fragments of heptapeptides of TM and NTM regions. The fragmentation was optimized based on Rosetta-CASP experiment[44], where maximum length of 7 residue fragments shows effective structure prediction in β-class proteins. (The fragments of different sizes, ranges from 5-9 residues do not affect the accuracy significantly) The closeness of the discriminant property value of the each heptapeptide fragment with the mean value of transmembrane and non-transmembrane regions was used to assign the location (TM/NTM). The accuracy was computed by comparing the assigned location with the experimentally derived location of the fragments using TMDET algorithm. This procedure was repeated for transmembrane/ non-transmembrane regions and for all superfamilies separately, and used to validate the efficiency of discrimination.

RESULTS AND DISCUSSION

Analysis on amino acid composition

The amino acid composition was calculated for the all membrane proteins of α and β class in the dataset (Fig 1a). Overall the hydrophobic residues Ala, Phe, Ile, Leu, Val are richer in α-class proteins, whereas the Asp, Gly, Ser, Thr are significantly higher in β class proteins; as expected, the polar residues are higher in the non-membrane regions of both classes. Interestingly, the amino acids predominantly found in α-class were due to their extensive involvement in the membrane bound regions. Overall a significant difference was observed in the amino acid composition between α and β class proteins: particularly Val, Leu, Ile, Phe, and sulphur containing residues (Cys, Met) are richer in α-class proteins, which correlates with earlier studies[45]. On the other hand the polar amino acids found dominant in β-structures (Asp, Ile, Ser, Thr), due to their extensive exposure in non-transmembrane regions as majority of the β structures were in β-barrel architecture, whereas the contribution of Ala, Leu and Gly were higher in its membrane bound regions (Fig. 1a). Further, we found that the amino acid composition of Ser, Asn and Gln are significantly dominant in β-class proteins which is consistent from the earlier study stating that these residues are important for the formation of β-barrel structures[31].

Role of amino acid composition in transmembrane and non-transmembrane segments

The analysis was further extended to study the amino acid composition at transmembrane and non-transmembrane regions based on their superfamily details (Table 2). Interestingly, the amino acid composition was observed as unique discriminative tool to identify the transmembrane/non-transmembrane segment of a protein of specific superfamily (Fig 1b). For example, in rhodopsin superfamily, the hydrophobic residues Phe, Ile, Leu, Gln and Val showed significant variation in the amino acid composition. Whereas, photosynthetic reaction centre and amino acid-polyamine-organocation superfamilies shows discriminative behaviour in the composition of Ala, Phe, Ile, Gly, Leu, Asn, Pro, Arg, Ser and Val; a similar trend was also observed in electron transport chain complex III and IV superfamily. Further the composition of Ile, Leu, Lys and Val are observed to be significantly varying between the transmembrane and non-transmembrane regions in all these superfamilies, whereas the polar residues Cys, Met, Trp and Tyr showed similar trend in all the superfamilies. Interestingly, there is no Asp, Glu and Asn amino acids found in the transmembrane regions of rhodopsin superfamily, however a slight contribution is observed in their non-transmembrane regions. In β-class protein structures, trimeric porins and autotransporters superfamily shows that there was a significant difference between the transmembrane regions and non-transmembrane regions in the Ala, Asp, Glu, Lys, Leu, Asn, Thr, Val and Tyr compositions.

In contrast, the polar residues His, Gln and Trp shows no significant variation between the transmembrane and non-transmembrane regions (Table 2).

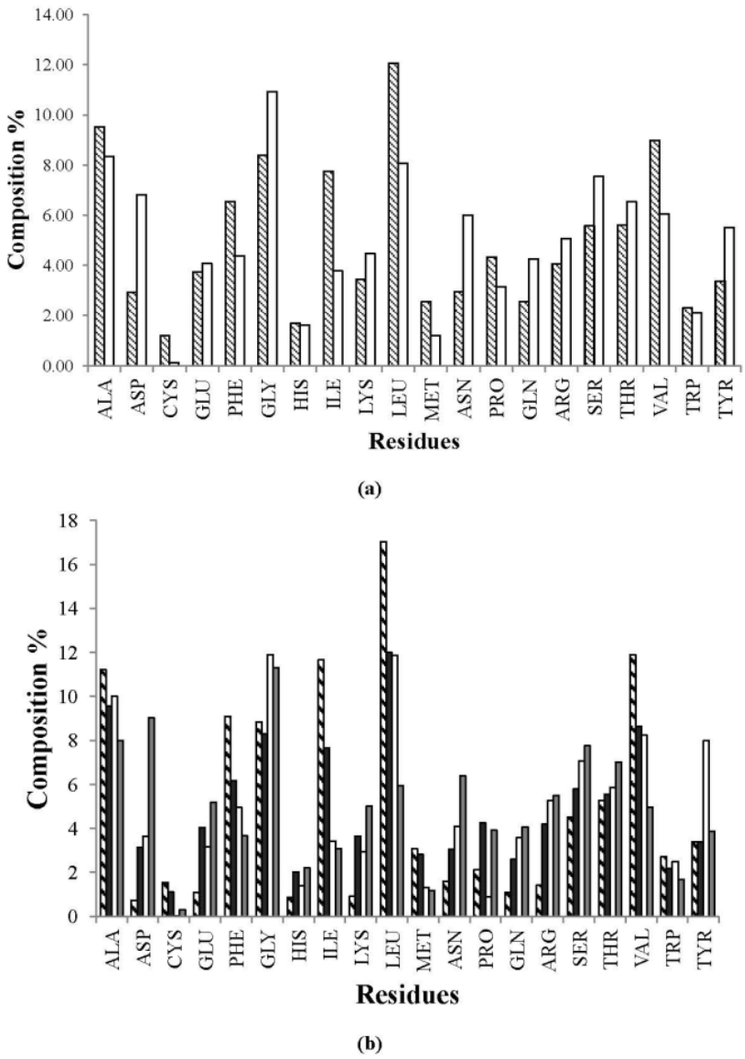

Figure 1

Amino acid composition in (a) α class [shaded] and β class [blank] membrane proteins and (b) transmembrane and non-transmembrane segments [shaded: α-class transmembrane, black: α-class non-transmembrane, white: β-class transmembrane and grey: β-class non-transmembrane]

Table 2

Amino acid composition of transmembrane and non-transmembrane segments of α and β-class proteins based on their superfamilies (TM- transmembrane, NTM- non-transmembrane regions)

	α-class [TM (NTM)]										β-class [TM (NTM)]			
	Rhodopsin	PRCP	APC	MFS	ETC III	MIP	Receptor	P type	Ion Channel	ETC IV	Ligand gated protein channels	Auto-Trans.	Trimeric proteins	oprD/algE
Ala	9.7(9.6)	13.6(9.4)	12.4(10.7)	11.9(10.8)	11.2(8.8)	13.3(11.5)	10.3(8.6)	11.0(7.9)	9.1(8.3)	12.0(10.0)	9.0(6.9)	11.6(10.2)	11.9(9.3)	8.6(6.8)
Asp	1.6(3.4)	0.1(2.9)	0.7(2.4)	0.6(2.6)	1.0(4.0)	0.3(2.4)	0.0(0.3)	0.8(4.9)	0.8(3.7)	0.3(3.2)	3.9(8.3)	3.8(9.4)	2.4(9.9)	4.6(9.2)
Cys	2.1(1.7)	1.3(0.6)	1.1(0.8)	1.8(1.0)	0.6(0.9)	2.2(1.5)	0.6(0.7)	2.7(1.8)	1.3(0.9)	1.0(0.6)	0.0(0.3)	0.0(0.0)	0.0(0.7)	0.1(0.0)
Glu	0.9(3.6)	0.4(3.6)	1.3(3.0)	1.0(2.8)	0.5(4.5)	1.7(2.6)	0.0(4.4)	2.2(6.4)	1.2(5.3)	1.5(4.2)	4.1(4.1)	3.0(5.2)	2.5(6.2)	1.9(6.0)
Phe	8.3(5.7)	11.6(7.2)	8.7(7.1)	8.8(7.1)	7.9(5.5)	8.5(6.7)	6.7(4.4)	8.5(4.2)	10.4(6.6)	10.0(6.6)	4.5(3.3)	3.5(3.8)	5.2(4.3)	6.9(3.6)
Gly	7.4(7.1)	9.5(7.9)	8.6(9.0)	11.3(9.5)	8.7(8.6)	11.6(11.2)	13.2(10.)	6.9(7.7)	7.7(6.9)	7.4(8.0)	10.5(9.4)	16.4(11.6)	12.8(12.2)	10.3(14.0)
His	0.8(2.3)	2.2(1.5)	0.5(1.2)	0.3(1.2)	0.6(2.7)	0.9(2.2)	0.6(1.3)	0.2(1.5)	0.4(2.1)	1.8(3.3)	1.2(2.1)	1.9(2.7)	1.2(1.5)	1.8(3.1)
Ile	11.1(7.3)	10.0(7.7)	11.8(9.6)	11.0(8.1)	7.5(5.3)	12.8(8.5)	15.1(9.8)	15.4(7.4)	11.7(7.7)	11.8(6.2)	3.6(3.9)	1.7(1.8)	4.2(3.0)	3.5(2.6)
Lys	1.1(3.6)	0.1(3.4)	1.0(3.1)	0.9(3.2)	2.2(4.6)	0.9(3.1)	0.6(4.4)	1.1(5.2)	0.9(3.2)	0.5(3.5)	3.3(4.0)	1.7(5.8)	3.1(6.8)	3.0(4.1)
Leu	16.6(12.3)	18.9(13.)	14.2(11.6)	14.7(12.4)	17.4(9.9)	14.6(11.2)	22.4(16.)	17.5(11.1)	18.2(12.1)	17.3(10.7)	11.9(6.1)	13.2(6.5)	9.2(4.8)	14.2(6.5)
Met	3.4(2.8)	2.9(2.9)	3.3(3.1)	4.3(3.8)	2.2(2.4)	3.0(2.5)	0.6(0.4)	2.3(2.6)	3.6(2.9)	3.4(3.8)	1.6(1.4)	1.2(1.3)	1.2(0.8)	0.9(1.2)
Asn	2.8(3.7)	1.0(3.3)	2.0(3.2)	2.1(3.0)	1.3(4.0)	1.1(3.6)	0.0(0.4)	2.0(3.1)	1.4(2.5)	0.9(2.9)	5.0(6.1)	2.2(6.3)	3.8(7.5)	4.0(5.6)
Pro	2.7(3.4)	2.4(5.1)	2.6(4.3)	2.8(4.1)	2.0(6.2)	2.0(3.7)	0.6(4.2)	2.0(3.9)	1.5(3.9)	1.8(5.0)	0.9(5.3)	1.5(2.1)	0.7(2.3)	0.7(4.5)
Gln	0.9(2.5)	0.7(3.1)	1.4(1.7)	2.3(2.9)	1.8(3.0)	1.4(1.8)	0.6(2.3)	1.2(2.8)	0.8(2.9)	0.6(3.0)	3.0(4.9)	3.3(3.4)	4.1(3.2)	4.4(3.9)
Arg	1.6(4.0)	1.4(4.0)	0.8(2.9)	1.2(3.5)	0.9(4.5)	0.7(2.2)	3.5(8.1)	1.3(5.4)	2.2(5.4)	0.7(3.1)	4.4(6.3)	6.7(6.9)	5.0(3.0)	6.3(5.9)
Ser	4.7(5.9)	3.2(5.1)	5.8(6.3)	5.9(6.4)	4.8(5.1)	4.4(6.2)	2.4(5.9)	4.6(5.8)	4.8(6.3)	3.9(4.7)	7.6(8.0)	6.3(7.7)	7.2(6.6)	6.3(9.0)
Thr	5.6(6.3)	3.2(4.8)	5.5(5.5)	4.3(4.7)	8.1(5.4)	5.5(6.0)	2.4(3.4)	4.9(5.5)	6.2(6.2)	5.5(5.5)	7.2(8.0)	4.1(6.1)	4.0(7.0)	6.8(5.8)
Val	10.9(8.4)	12.2(8.8)	11.2(8.6)	9.5(7.8)	13.0(7.8)	10.3(8.2)	18.6(12.)	11.2(8.4)	12.5(8.5)	12.4(8.8)	8.0(5.3)	8.6(3.7)	10.5(6.2)	5.4(3.8)
Trp	3.2(2.3)	3.4(2.6)	3.3(2.5)	2.3(2.2)	4.1(2.7)	2.2(2.1)	0.6(0.6)	1.4(1.3)	1.3(1.7)	4.6(3.3)	2.8(1.6)	2.3(1.8)	1.9(2.0)	2.8(1.3)
Tyr	4.9(4.2)	2.1(3.1)	3.8(3.5)	3.2(3.0)	4.6(4.2)	2.5(2.8)	1.2(1.4)	2.8(3.4)	3.9(3.1)	2.8(3.8)	7.6(4.8)	7.0(3.7)	9.4(3.0)	7.8(3.2)

Amino acid properties in the transmembrane and non-transmembrane regions

The amino acid properties are calculated for the overall proteins in the dataset and also separately for transmembrane and non-transmembrane regions. We found that 17 and 19 properties were significant in the transmembrane regions of α and β class respectively, whereas the remaining are significant in non-transmembrane regions. Since significance on either side (TM or NTM) has the useful information to discriminate the transmembrane and non-transmembrane segments of proteins, we considered both the states. Particularly the results show that, in α-class proteins, the properties, thermodynamic transfer hydrophobicity, short and medium range non-bonded energy, total non-bonded energy, long-range non-bonded energy etc. are significant to identify the transmembrane regions (Table 3), whereas properties such as turn tendency, coil tendency, backbone dihedral probability etc., are useful to identify the non-transmembrane segments. Interestingly the physico-chemical properties such as thermodynamic transfer hydrophobicity, short, medium and long range non-bonded energies, total non-bonded energy, average number of surrounding residues, average medium and long range contacts, Gibbs free energy change and unfolding hydration heat capacity change are appeared to be common significant properties to identify the transmembrane regions in both the α and β class proteins. In addition, solvent accessible surface area for denatured protein and shape (position of branch point in a side-chain) are observed to have significance in the transmembrane regions of β class proteins (Table 3).

Significance of discriminant properties using T-test

The analysis on T-test confirms that many of the properties are significant to discriminate transmembrane and non-transmembrane regions (Table 4). Particularly in α-class, total non-bonded energy, average long-range contacts, average number of surrounding residues, Gibbs free energy change of hydration and hydration for denatured protein are found most significant (100%) for all α-class proteins. Together with the total non-bonded energy and average number of surrounding residues, unfolding hydration heat capacity change are also found most significant for all β-class proteins. In addition, long-range non-bonded energy, Gibbs free energy change of hydration for native protein, beta-helical tendency, short and medium range non-bonded energy, thermodynamic transfer hydrophobicity, unfolding enthalpy change are found capable in discriminating α-class proteins with 90-99% significance; of these, long-range non-bonded energy, unfolding enthalpy change, Gibbs free energy change of hydration for native protein and thermodynamic transfer hydrophobicity are observed to be potential properties to discriminate β- class proteins as well with 90-99% significance. Finally, the properties such as average medium-range contacts, unfolding entropy change of hydration, unfolding enthalpy change of chain can discriminate transmembrane regions in α-class proteins with the minimum significant range (50-90%). Similarly, short and medium range non-bonded energy, unfolding enthalpy change of hydration, solvent accessible surface area for denatured protein are able to discriminate transmembrane regions of β-class. In contrast, unfolding enthalpy change of chain and equilibrium constant with reference to the ionization property of COOH group (pK′) shows the fluctuation behaviour in discriminating the α-class proteins due to the poor assessment in ion channel, p-type ATPase and major intrinsic protein superfamilies. Similarly, short and medium range non-bonded energy, solvent accessible surface area for denatured protein and unfolding enthalpy change of hydration are found less significant in β-class due to trimeric porins and oprD/algE superfamilies.

Table 3
Mean values of the amino acid properties for transmembrane (TM), non-transmembrane regions (NTM) and in overall protein of α- and β-class

Properties‡	α-class			β-class		
	TM	NTM	Overall	TM	NTM	Overall
Ht	44.3	38.6	39.2	34.1	28.4	36.1
pK'	72.6	73.7	74.2	79.6	75.5	68.5
Esm	67.7	64.1	64.6	63.2	61.3	62
El	58.4	49.8	50.9	48.5	39.6	49
Et	73.5	63.8	64.9	62	53.5	61.5
Pb	61.1	52.7	54	52.9	43.7	50.5
Pt	30.5	40.4	40.1	45.9	54.6	44.9
Pc	27.2	35.4	35.1	38.8	46.2	36.8
F	41	53.2	52.6	58.5	67.4	54.9
Ns	58.8	48.2	49.4	46.4	35.6	45.2
aC	26.8	32.3	32.2	30.1	38	35.9
Nm	55.7	54.9	55.1	54.4	51.7	58
Nl	61.8	52.2	53.5	51.6	42	51
ASAD	52.3	53.9	54.2	52.8	50.9	58.2
ASAN	21.5	31.3	30.7	32.3	39	33.6
dGh	77	69.4	70.1	63.1	60	59.1
GhD	81.4	73.9	74.8	69.3	65.2	66.8
GhN	85.9	79.5	80.5	76.3	72	74.8
dHh	69.9	66.8	67.3	63	63.9	59.8
-TdSh	49.9	46.6	47.3	46.5	39.8	47.4
dCph	58.8	48.8	49.9	44.4	34.1	42.3
dGc	22.9	29.8	29.7	36.2	37.6	40.5
dHc	28.8	28.5	29	31.2	27.6	36
-TdSc	56.6	64	63.7	65.8	72.7	61.8
dG	32.1	28.8	29.3	28.7	25.3	30.5
dH	47.3	43.7	44.4	43.3	39.9	46.4
-TdS	45.1	50.4	50.1	51.8	55.2	48.3
f	28	32.5	32.3	27.4	27.3	25.6
s	22	25	25.1	32.6	34.3	36.6
Pf-s	15.4	20.3	20	24.7	29.3	22.5

‡Ht: Thermodynamic transfer hydrophobicity; pK': Equilibrium constant with reference to the ionization property of COOH group; Esm: Short and medium range non-bonded energy; El: Long-range non-bonded energy; Et: Total non-bonded energy; Pb: Beta-helical tendency; Pt: Turn tendency; Pc: Coil tendency; F:Mean rms fluctuational displacement; Ns: average number of surrounding residues; aC: Power to be at the C-terminal of alpha helix; Nm: Average medium-range contacts; Nl: Average long-range contacts; ASAD: Solvent accessible surface area for denatured protein; ASAN:Solvent accessible surface area for native protein; dGh: Gibbs free energy change of hydration for unfolding; GhD: Gibbs free energy change of hydration for denatured protein; GhN: Gibbs free energy change of hydration for native protein; dHh: Unfolding enthalpy change of hydration; -TdSh: Unfolding entropy change of hydration; dCph: Unfolding hydration heat capacity change; dGc: Unfolding Gibbs free energy change of chain; dHc: Unfolding enthalpy change of chain; -TdSc: Unfolding entropy change of chain; dG: Unfolding Gibbs free energy change; dH: Unfolding enthalpy change; –TdS: Unfolding entropy change; s:Shape (position of branch point in a side-chain; f:Flexibility (number of side-chain dihedral angles; Pf-s:Backbone dihedral probability.

Table *4*
Significance of properties based on T-test analysis in α-class and β-class proteins. The abbreviations of the properties are given in Table3.

α-class proteins				β-class proteins			
Properties	Significance (from T-test)	Location	%	Properties	Significance (from T-test)	Location	%
Et	Y	TM	100	Et	Y	TM	100
Nl	Y	TM	100	Ns	Y	TM	100
Ns	Y	TM	100	dCph	Y	TM	100
dGh	Y	TM	100	El	Y	TM	98.5
GhD	Y	TM	100	Nl	Y	TM	98.5
El	Y	TM	99.18	dH	Y	TM	98.5
GhN	Y	TM	99.32	GhD	Y	TM	96.4
Pb	Y	TM	98.97	GhN	Y	TM	94.28
dCph	Y	TM	97.95	dG	Y	TM	93
Esm	Y	TM	95.21	Ht	Y	TM	92.85
dH	Y	TM	95.21	pK'	Y	TM	90.0
Ht	Y	TM	94.18	-TdSh	Y	TM	87.14
dHh	Y	TM	94.18	dGh	Y	TM	84.28
dG	Y	TM	91.44	Nm	Y	TM	81.81
-TdSh	Y	TM	89.38	dHC	Y	TM	80.0
Nm	Y	TM	75.83	Pa	N	TM	78.78
dHc	N	-	55.82	Esm	Y	TM	69.43
pK'	N	-	50.68	dHh	N	-	69.69
ASAD	Y	NTM	34.25	ASAD	N	-	69.69
s	Y	NTM	13.70	s	N	-	42.85
f	Y	NTM	7.19	dGc	Y	NTM	31.42
aC	Y	NTM	5.82	F	Y	NTM	25.70
Pf-s	Y	NTM	5.82	Pc	Y	NTM	11.42
-TdSc	Y	NTM	5.14	Pt	Y	NTM	10.0
-TdS	Y	NTM	3.77	pf-s	Y	NTM	10.0
Pc	Y	NTM	3.42	-TdSc	Y	NTM	8.50
dGc	Y	NTM	2.74	-TdS	Y	NTM	5.71
Pt	Y	NTM	1.37	F	Y	NTM	5.71
F	Y	NTM	0.68	aC	Y	NTM	2.84
ASAN	Y	NTM	0.34	ASAN	Y	NTM	2.87

Discriminant properties at the superfamily level

The efficiency of significant properties (computed from T-test at class level) to discriminate the transmembrane regions are further examined at the superfamily level, using the larger deviation of those significant property values between transmembrane and non-transmembrane regions as a probe. We have performed the test within the superfamily, and selected top four significant properties and consider them as potential discriminant properties. Interestingly, average number of surrounding residues, total non-bonded energy and average long-range contacts are observed as the potential discriminate properties in majority of the superfamilies in α-class proteins (Fig. 2a). Along with these, unfolding hydration heat capacity change and long-rage non-bonded energy are observed as potential properties in β-class proteins (Fig 2b). Interestingly, our result emphasis the importance of considering the number of residues surrounding a particular residue within the effective distance of influence to identify transmembrane segment as it changes the surrounding hydrophobicity indices of the amino acids. Similarly the total non-bonded energy and long-range contacts, the indicators of protein stability and folding[46], should also be a potential component to discriminate the transmembrane and non-transmembrane regions.

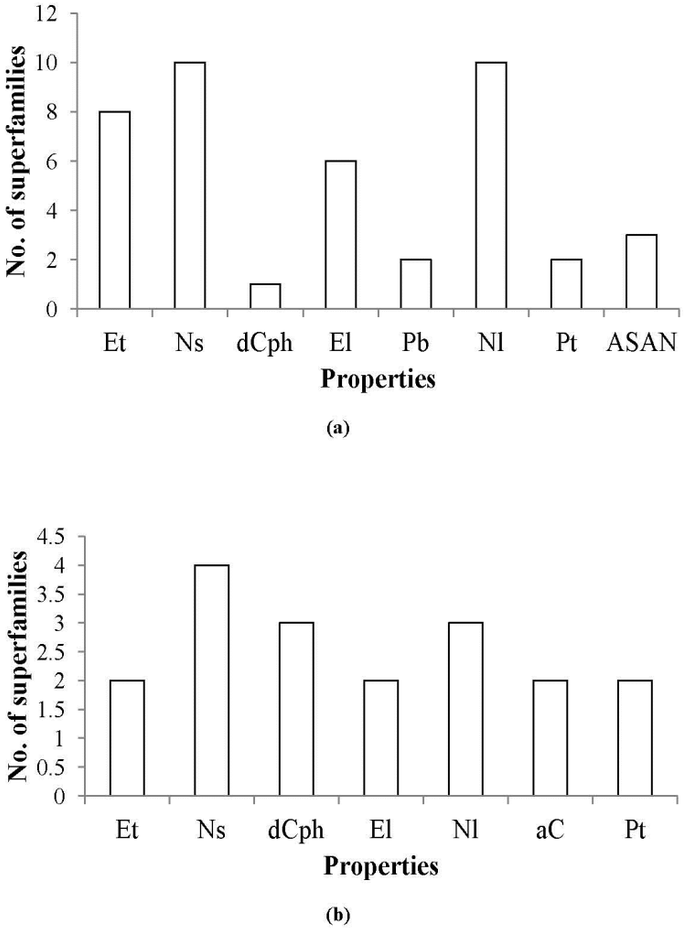

Figure 2
Discriminant properties for no. of superfamilies in (a) α-class proteins and (b) β-class proteins. The abbreviations of the properties are given in Table3.

Validation

The accuracy of discrimination is computed using heptapeptide fragmentation approach and our results demonstrates that the amino acid properties are efficient probes to discriminate the membrane proteins of both α and β-class (Fig 3). Overall the amino acid properties discriminates the α-class superfamily proteins with an average accuracy of 77%. Particularly, transmembrane segments of p-type ATPase and photosynthetic reaction centre superfamily proteins are discriminated with the accuracy of 84.8% and 82.4% respectively. Similarly the β-class proteins discriminates the transmembrane segments with an of 69.7%; of these trimeric porins superfamily proteins are discriminated with higher accuracy of average accuracy 75%.

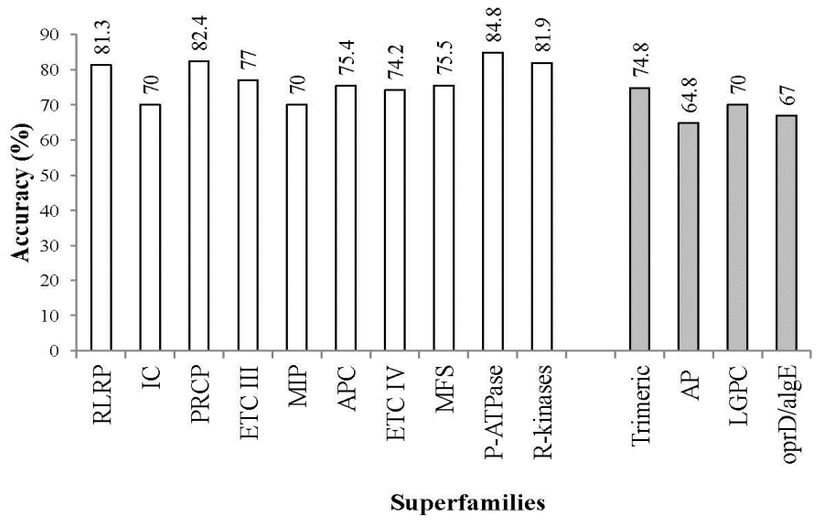

Figure 3

Accuracy of discrimination (in %) at the superfamily level in α and β-class proteins. RLRP: Rhodopsin-like receptors and pumps; IC: Ion channel super family; PRCP: Photosynthetic reaction centres and photosystems; ETC III: Electron transport chain complex III; MIP: Major intrinsic protein; ETC IV: Electron transport chain complex III; MFS: Major facilitator superfamily; P-ATPase: P type ATPase; R-kinases: Receptor type kinases; AP: Autotransporters; LGPC: Ligand gated protein channels (Blank: α-class superfamilies, Shaded: β-class superfamilies)

CONCLUSION

An improved method to predict the transmembrane segments of a membrane protein from the amino acid sequence is an important need in biology and this can be achieved by considering various biophysical and biochemical factors of the membrane regions. Studies have utilized different strategies towards the accurate prediction and many of them show better performance for the specific class of proteins[6, 28,34,38]. Herein we performed a simple systematic analysis which emphasis the importance of various physico-chemical amino acid properties to discriminate the transmembrane regions particularly at the class and superfamily level. We found that several properties can do this work efficiently: particularly, information about the average number of surrounding residues, total non-bonded energy of transmembrane region, long-range contacts, Gibbs free energy change of hydration, unfolding enthalpy change, and thermodynamic transfer hydrophobicity clearly discriminate membrane bound regions of both α and β class proteins. Of these, average numbers of surrounding residues, average long-range contacts and total non-bonded energy are capable to identify transmembrane segments of proteins of all superfamilies with the accuracy of 91, 89 and 85% respectively. Our results emphasis the importance of considering the amino acid properties and its discrimination capacity at the superfamily level; the accuracy values are comparable with other prediction strategies. Based on the results, we suggest that the integration of these parameters along with the existing strategies can help to improve the accuracy of the prediction. However, further computational studies are required in a larger training and test datasets, to understand the specific performance of these properties in transmembrane region discrimination.

ACKNOWLEDGEMENT

This work was done as a part of the course work for the award of M.Tech.-Bioinformatics degree to C. Vignesh. The authors wish to thank the management of SASTRA University for the support and encouragement.

CONFLICT OF INTEREST

Conflict of interest declared none.

REFERENCES

1. Punta M., Forrest LR., Bigelow H., Kernytsky A., Liu J., Rost B. Membrane protein prediction methods. Methods, 41 (4): 460-474, (2007)
2. Almén MS., Nordström KJ., Fredriksson R., Schiöth HB. Mapping the human membrane proteome: A majority of the human membrane proteins can be classified according to function and evolutionary origin. BMC Biology, 7 : 50, (2009)
3. Sachs JN., Engelman DM. Introduction to the membrane protein reviews: the interplay of structure, dynamics, and environment in membrane protein function. Annu. Rev. Biochem., 75 : 707-712, (2006)
4. Marsico A., Labudde D., Sapra T., Muller DJ., Schroeder M. A novel pattern recognition algorithm to classify membrane protein unfolding pathways with high-throughput single-molecule force spectroscopy. Bioinformatics, 23 (2): 231-236, (2007)
5. Arunya J. A review on a new tetraspanin, TSPAN-11. Int. J. Pharm. Bio Sci., 3 (3): 418-426, (2012)
6. Wang H., Zhang C., Shi X., Zhang L., Zhou Y. Improving transmembrane protein consensus topology prediction using inter-helical interaction. Biochim. Biophys. Acta, 1818 (11): 2679-2686, (2012)
7. Freeman TC Jr., Wimley WC. A highly accurate statistical approach for the prediction of transmembrane β-barrels. Bioinformatics, 26 (16): 1965-1974, (2010)
8. Houck SA., Cyr DM. Mechanisms for quality control of misfolded transmembrane proteins. Biochim. Biophys. Acta, 1818 (4): 1108-1114, (2012)
9. Bagos PG., Liakopoulos TD., Spyropouls IC., Hamodrakas SJ. A Hidden Markov Model method, capable of predicting and discriminating β-barrel outer membrane proteins. BMC Bioinformatics, 5 : 29, (2004)
10. Tusnady GE., Dosztanyi Z., Simon I. Transmembrane proteins in the Protein Data Bank: identification and classification. Bioinformatics, 20 (17): 2964-2972, (2004)
11. Tsaousis GN., Bagos PG., Hamodrakas SJ. HMMpTM: Improving transmembrane protein topology prediction using phosphorylation and glycosylation site prediction. Biochem. Biophys. Acta, 1844 (2): 316-322, (2014)
12. Simakova MN., Simakov NN., Computational Methods for predicting Structure of Membrane Proteins Using Amino Acid Sequences. Molecular Biology, 47 (2): 307-315, (2012)
13. von Heijne G. Membrane protein structure prediction – hydrophobicity analysis and the positive inside rule. J. Mol. Biol., 225 (2): 487-494, (1992)
14. Tusnady GE, and Simon I, The HMMTOP transmembrane topology prediction server. Bioinformatics, 17 (9): 849-850, (2001)
15. Ganapathiraju M., Balakrishnan N., Reddy R., Seetharaman JK. Transmembrane helix prediction using amino acid property features and latent semantic analysis. BMC Bioinformatics, 9 (Suppl 1): 1-16, (2008)
16. Ramanathan K., Shanthi V., Sethumadhavan R. Exploring the role of C–H….π interactions on the structural stability of membrane proteins. Int. J. Pharm. Bio Sci., 1 (1): 1-14, (2010)
17. Jones DT. Do transmembrane protein superfolds exist? FEBS Lett., 423 (3): 281-285, (1998)
18. Kitsas IK., Hadjileontiadis LJ., Panas SM. Transmembrane helix prediction in proteins using hydrophobicity properties and higher-order statistics. Comput. Biol. Med., 38 (8): 867-880, (2008)
19. Koehler J., Woetzel N., Staritzbichler R., Sanders CR., Meiler J. A unified hydrophobicity scale for multispan membrane proteins. Proteins, 76 (1): 13-29, (2009)
20. Rees DC., DeAntonio L., Eisenberg D. Hydrophobic organization of membrane proteins. Science, 245 (4917): 510-513, (1989)
21. Kihara D., Shimizu T., Kanehisa M. Prediction of membrane proteins based on classification of transmembrane segments. Protein Eng., 11 (11): 961-970, (1998)
22. Claros MG., von Heijne G. TopPred II: an improved software for membrane protein structure predictions. Comput Appl Biosci, 10 (6): 685-686, (1994)
23. Hirokawa T., Boon-Chieng S., Mitaku S. SOSUI: classification and secondary structure prediction system for membrane proteins. Bioinformatics, 14 (4): 378-379, (1998)
24. Jones DT., Taylor WR., Thornton JM. A model recognition approach to the prediction of all-helical membrane protein structure and topology. Biochemistry, 33 (10): 3038-3049, (1994)

25. Melen K., Krogh A., von Heijne G. Reliability Measures for Membrane Protein Topology Prediction Algorithms. J. Mol. Biol., 327 (3): 735-744, (2003)
26. Rost B., Cassadio R., Fariselli P. Refining neural network predictions for helical transmembrane proteins by dynamic programming. Proc. Int. Conf. Intell. Syst. Mol. Biol., 4 : 192-200, (1996)
27. Fuchs A., Martin-Galiano AJ., Kalman M., Fleishman S., Ben-Tal N., Frishman D. Co-evolving residues in membrane proteins. Bioinformatics, 23 (24): 3312-3319, (2007)
28. Zou L., Wang Z., Wang Y., Hu F. Combined prediction of transmembrane topology and signal peptide of beta-barrel proteins: using a hidden Markov model and genetic algorithms. Comput. Biol. Med., 40 (7): 621-628, (2010)
29. Zheng WJ., Spassov VZ., Yan L., Flook PK., Szalma S. A hidden Markov model with molecular mechanisms energy-scoring function for transmembrane helix prediction. Comput. Biol. Chem., 28 (4): 265-274, (2004)
30. Martelli PL., Fariselli P., Krogh A., Casadio R. A sequence-profile-based HMM for predicting and discriminating beta barrel membrane proteins. Bioinformatics, 18 (Suppl 1): S46-S53, (2002)
31. Gromiha MM., Ahmad S., Suwa M. TMBETA-NET: discrimination and prediction of membrane spanning β-strands in outer membrane proteins. Nucleic Acids Res., 33 : 164-167, (2005)
32. Yuan Z., Davis MJ., Zhang F., Teasdale RD. Computational differentiation of N-terminal signal peptides and transmembrane helices. Biochem. Biophys. Res. Commun., 312 (4): 1278-1283, (2003)
33. Jones DT. Improving the accuracy of transmembrane protein topology prediction using evolutionary information. Bioinformatics, 23 (5): 538-544, (2007)
34. Kall L., Krogh A., Sonnhammer EL. A combined Transmembrane Topology and Signal Peptide Prediction method. J. Mol. Biol., 338 (5): 1027-1036, (2004)
35. Nugent T., Jones DT. Transmembrane protein topology prediction using support vector machines. BMC Bioinformatics, 10 : 159, (2009)
36. Moult J., Fidelis K., Kryshtafovych A., Rost B, Tramontano A. Critical assessment of methods of protein structure prediction – Round VIII. Proteins, 77 (Suppl 9): 1-4, (2009)
37. Kelm S., Shi J., Deane CM. MEDELLER: homology-based coordinate generation for membrane proteins. Bioinformatics, 26 (22): 2833-2840, (2010)
38. Gromiha MM., Suwa M. Influence of amino acid properties for discriminating outer membrane proteins at better accuracy. Biochim. Biophys. Acta, 1764 (9): 1493-1497, (2006)
39. Nugent T., Jones DT. Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis. Proc. Natl. Acad. Sci., 109 (24): 1540-1547, (2012)
40. Elofsson A., von Heijne G. Membrane protein structure: prediction versus reality. Annu. Rev. Biochem., 76 : 125-140, (2007)
41. Kozma D., Simon I., Tusnady GE. PDBTM: Protein Data Bank of transmembrane proteins after 8 years. Nucleic Acids Res., 41 : 524-529, (2013)
42. Tusnady GE., Dosztanyi Z., Simon I. TMDET: web server for detecting transmembrane regions of proteins by using their 3D coordinates. Bioinformatics, 21 (7): 1276-1277, (2004)
43. Gromiha MM., Oobatake M., Sarai A. Important amino acid properties for enhanced thermostability from mesophilic to thermophilic proteins. Biophys. Chem., 82 (1): 51-67, (1999)
44. Raman S., Vernon R., Thompson J., Tyka M., Sadreyev R., Pei J., Kim D., Kellogg E., DiMaio F., Lange O., Kinch L., Sheffler W., Kim BH., Das R., Grishin NV., Baker D. Structure prediction for CASP8 with all-atom refinement using Rosetta. Proteins, 77 (Suppl 9): 89-99, (2009)
45. Gromiha MM., Suwa M. A simple statistical method for discriminating outer membrane proteins with better accuracy. Bioinformatics, 21 (7): 961-968, (2005)
46. Gromiha MM., Selvaraj S. Importance of long-range interactions in protein folding. Biophys. Chem., 77 (1): 49-68, (1999).