Method for Predicting Transmembrane Helices in Protein Sequences

Raazia Rahim (PG Scholar)
Department of computer Science and Engineering
Younus college of Engineering and Technology
Pallimukku , Kollam ,691010

Prof Nijil Raj N
Head Of The Department
Department of computer Science and Engineering
Younus college of Engineering and Technology
Pallimukku , Kollam ,691010

***Abstract*—The increasing protein sequences from the genome project require the oretical methods to predict transmembrane helical segments (TMHs). So far, several prediction methods have been reported, but there are some deficiencies in prediction accuracy and adaptability in these methods. Here, a method based on discrete wavelet transform (DWT) has been developed to predict the number and location of TMHs in membrane proteins,80 proteins with known 3D structure from Mptopo database are chosen at random as data sets (including 325 TMHs).TMHs prediction is carried out for the membrane protein sequences and obtain satisfactory result. To verify the feasibility of this method, 80 membrane protein sequences are treated as test sets, 308 TMHs can be predicted and the prediction accuracy is 96.3%Compared with the other prediction results , the obtained results indicate that the proposed method has higher prediction accuracy.**

***Index Terms*—Membrane Protein, Transmembrane Helices,Hydrophobicity.**

I. Introduction

A transmembrane protein (TP),which is a type of integral membrane protein that spans the entirety of the biological membrane to which it is permanently attached. Many transmembrane proteins function as gateways to allow the transport of specific substances across the biological membrane. They undergo significant changes to move a substance through the membrane.

Transmembrane proteins are polytopic proteins that aggregate and precipitate in water. They require detergents or nonpolar solvents for extraction, although some of them (beta-barrels) can be also extracted using denaturing agents.

The other type of integral membrane protein is the integral monotopic protein that is also permanently attached to the cell membrane but does not pass through it.

The knowledge of the function of membrane protein itself has been expanded enormously and deeply, and the more study of it can be used as a breakthrough of studying protein structure and function and the genetic information in DNA sequence. In order to explore the relationship between membrane protein structure and function, understand various work mechanism in membrane protein life activities, bioinformatics methods and techniques of developing the study of membrane protein are needed. In the genome data, a large portion (about 20%-30%) of proteins in a genome encodes membrane protein [1-3] , the proportion of such shows the importance of membrane protein in biology. Membrane protein, especially transmembrane protein has very important function in organism, such as photosynthesis, respiration, neural signaling, immune response, nutrient absorption and so on, and it is also the important drug target. Of the drug target known and being researched is about 70% of the membrane protein [4]. Here introduces a method for predicting the Transmembrane Helices(TMH) in a protein sequences. The introduced method make use of hydrophobocity values of the aminoacid sequences and a wavelet transform.We are hoping better prediction accuracy for the method when compared to all other existing method

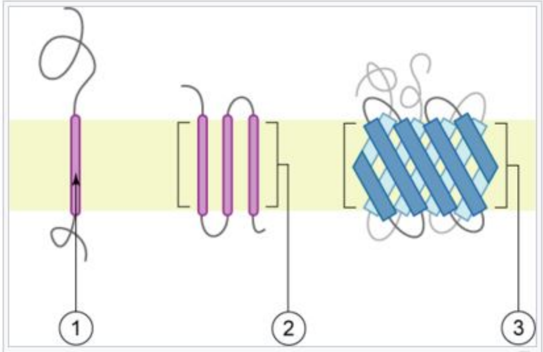


Fig. 1. Schematic representation of transmembrane

II. Literature Survey

Many transmembrane helical segments (TMHs) predicting algorithms for membrane proteins have been proposed. In 1982 Kyte and Doolittle firstly suggested a hydrophobicity analysis method of membrane protein sequences [5]. Thereafter von Heijne putforward the well-known ”positive-inside rule” to guide prediction in 1986 [6]. SOSUI [7], PRED-TMR [8] were based on the foregoing two methods.

In recent years, some statistical methods have been developed that like DAS [9], TMAP [10332], neural networks PHDhtm [10, 11], TMHMM [1, 12] and HMMTOP [13,14] based on hidden Markov model, MEMSAT-SVM prediction method based on support vector machine [15, 16].

Wavelet transform was first introduced into bioinformatics research in 1996 [17] and raised extensive attention immediately [18. Lio et al [19] proposed a non-parametric method based on a wavelet data-dependent threshold technique for change-point analysis which was applied to predict TMHs in membrane proteins. Continuous wavelet transform (CWT) for predicting the number and location of helices in membrane proteins is presented by Qiu et al. Pashou et al applied a dynamic programming algorithm on wavelet-denoised 'hydropathy' signals to determine membrane spanning segments.

Here we make full use of the hydrophobicity of amino acids and multiresolution feature of discrete wavelet transform (DWT) to decompose the amino acids of TM proteins into a series of structures in different layers, then predicting the location of TMHs according to the information of the amino acids sequence in different scales.

III. Materials and Method

A. Materials

The test data set is collected from the MPtopo database, which consist of a set of membrane protein structure which can be treated as reliable samples.The test dataset consist of 80 protein sequences with known 3D structure.The data can be obtained from http://blanco.biomol.uci.edu/mptopo.[23]

B. Method

The feature of protein structure is the balance between hydrophobic and hydrophilic and the structure stability depends heavily on molecules hydrophobic effects [20-21]. The determination of hydrophobic value of amino acid is mainly calculated according to distribution coefficient in which various amino acid . So when we map the amino acid sequence of protein onto a sequence of hydrophobicity, we need to optimize a variety of different hydrophobic parameters. Here, we use KD hydrophobic parameter values.

In order to predict TMHs of membrane protein sequence, with the condition of selecting the suitable wavelet basis functions and threshold are important.The threshold here is determined by the maximum average prediction accuracy of training set. Using this threshold, we are able to predict TMHs among membrane protein sequences from test set.

Procedure is as follows:

1) According to their hydrophobic amino acid value, convert 80 amino acid sequence of membrane protein into a sequence of hydrophobicity value.

TABLE I
Membrane protein family used for prediction

Family Name	PDB Code		
Bacteriorhodopsin	1ap9		
ABC transporters	1jsq	1l7vA	1pf4
Channel proteins	1fqyA	1fx8A	1msl
	1mxm	1oedA	1oedB
	1oedC	1oedE	1p7b
	1rc2A	1rhzA	1rhzB
Cytochrome bc1 complexes	1bgyE	1bgyJ	1bgyK
Cytochrome b6f complexes	1um3A	1um3B	1um3D
	1um3F	1um3G	1um3H
Cytochrome c oxidases	1ehkA	1ehkB	1ehkC
	1occA	1occB	1occC
	1occD	1occG	1occI
	1occJ	1occK	1occL
	1occM	1qleA	1qleB
	1qleC	1qleD	
Glycophorin	1afoA		
Light-harvesting complexes	1kzuA	1lghA	
Photosynthetic reaction centers	1eysH	1eysL	1eysM
	1prcH	1prcL	1prcM
	2rcrL	2rcrM	
Photosystems	1jboA	1jboB	1jboF
	1jboI	1jboJ	1jboK
Respiratory proteins	1a91C	1fftA	1fftB
	1fftC	1fumC	1kqgB
	1kqgC	1lovD	1nekC
	1nekD	1okcA	1q16C
	1qlaC		
Rhodopsins	1f88	1h2sB	1h68A
Translocation proteins	1pw4A	1s7b	2cpb

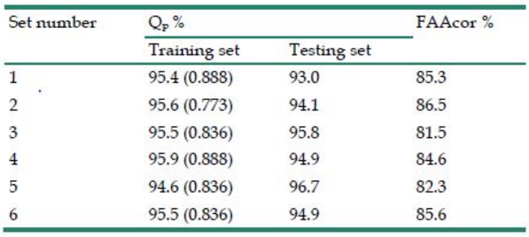
2) Randomly choose the training set from the 80 protein sequences and rest of them are considered as test data.
3) According to the data of training set, analyze and determine wavelet function.
4) Discrete Wavelet Transform (CWT) is used to find out the wavelet coeffient and the optimized threshold.
5) Get prediction result by predicting samples of the test set, and do statistics and analysis of the precision of prediction compared with experimental data.

This method is implementing in MATLAB for the convienence.

C. Evaluation of Result

Due to the limitations of experimental condition, we hope that the predicted TMHs are regarded as correct when over half of the predicted TMHs coincide with the observed TMHs. In statistical analysis, the average length of TMHs is 20 a.a. In this method, we decide that predicted TMHs are correct when at least 9 continuous residues are contained in the observed TMHs. The prediction accuracy of TMHs, $Q_p = \sqrt{M * C} * 100$, where M=Ncor/Nobs (Ncor stands for the number of correctly predicted TMHs, Nobs stands for the number of observed TMHs);M can be regard as a measure index of sensitivity; C=Ncor/Nprd (Nprd stands for the total number of predicted TMHs), C is regarded as a measure index of specificity.
For sequence of KD hydrophobic parameters, db10 is used as optimal wavelet basis. At scale level j=4, data of each group

of training set obtained the corresponding optimal threshold. In the test set, we use the threshold 0.836, get maximum average prediction accuracy of the membrane protein TMHs being 95.8%, maximum average prediction accuracy of residue being 83.1%.



Set number	Q_p %		FAAcor %
	Training set	Testing set	
1	95.4 (0.888)	93.0	85.3
2	95.6 (0.773)	94.1	86.5
3	95.5 (0.836)	95.8	81.5
4	95.9 (0.888)	94.9	84.6
5	94.6 (0.836)	96.7	82.3
6	95.5 (0.836)	94.9	85.6

Fig. 3. Prediction accuracy for each group of training set and test set of KD hydrophobic parameters.

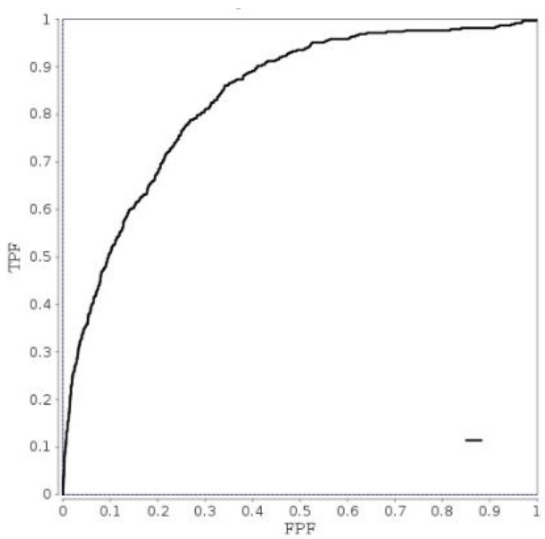


Fig. 4. ROC curve of the proposed method

IV. Conclusion

Transmembrane proteins are polytopic proteins that aggregate and precipitate in water. They require detergents or nonpolar solvents for extraction, although some of them (beta-barrels) can be also extracted using denaturing agents. Here introduces a method for predicting the Transmembrane Helices(TMH) in a protein sequences. The introduced method make use of hydrophobocity values of the aminoacid sequences and a wavelet transform.We are hoping better prediction accuracy for the method when compared to all other existing method.

Although the proposed method has the characteristics of simplicity, visual process, and high accuracy, through the analysis of the predicted results of data set, it is found that compared with the actual structure of membrane protein, there still exist some differences in the position and number of the predicted TMHs. They are as follows: too much prediction, that is the position and number of the predicted TMHs and the actual structure is not completely corresponding; less prediction, i.e. it haven't predicted all the position and number of the actual structure of TMHs. This is because: (1) When using wavelet transform, we just map the amino acid sequence of membrane protein into hydrophobic value sequence.The hydrophobic effect is the most important factor to determine the stability of protein structure, it is not the only factor. Beyond the hydrophobic effect, there are hydrogen bond, ionic bond and van der Waals force and disulfide bond of peptide chain, etc; (2) The volume of protein molecules, electric charge and many kinds of factors all have the regulation effects on the protein structure and stability; (3) Based on the signal peptide hypothesis, the signal peptide can form TMHs in protein synthesis, auxiliary peptide chain across the endoplasmic reticulum (ER) membrane, so in forecasting TMHs, it is very normal that signal peptide is contained. Because hydrophobicity is the main sequence characteristic of transmembrane helices, and there are likely to be long hydrophobic sequence in the hydrophobic core of water-soluble globular protein, which also can produce false positive results.When considering many kinds of factors, and when predicting the position and number of membrane protein TMHs by mapping the amino acid sequence into hydrophobic value sequence, the deviation within the scope is .allowed. If the above many factors are considered, the prediction accuracy can be improved.

References

[1] Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol. 2001; 305: 567-580, .

[2] Liu J, Rost B. Comparing function and structure between entire proteomes. Protein Sci. 2001; 10: 1970-1979.

[3] Kihara D, Shimizu T, Kanehisa M. Prediction of membrane proteins based on classification of transmembrane segments. Protein Engin. 1998; 11: 961-970.

[4] Yildirim MA, Goh KI, Cusick ME, Barabasi AL, Vidal M. Drug-target network. Nature Biotechnology. 2007; 25: 1119-1126.

[5] Kyte J, Doolittle RF. A simple method for displaying the hydrophathic character of a protein. J Mol Biol. 1982; 157: 105-132.

[6] von Heijne G. The distribution of positively charged residues in bacterial inner membrane proteins correlates with the transmembrane topology. EMBO J. 1986; 5: 3021-3027.

[7] Hirokawa T, Boon-Chieng S, Mitaku S. SOSUI: classification and secondary structure prediction system for membrane proteins. Bioinformatics. 1998; 14: 378-379.

[8] Pasquier C, Promponas VJ, Palaios GA, Hamodrakas JS, Hamodrakas SJ. A novel method for predicting trsnsmembrane segments in proteins based on a statistical analysis of the SwissProt database: the PRED-TMR algorithm. Protein Eng. 1999; 12: 381-385.

[9] Cserz M, Wallin E, Simon I, von Heijne G, Elofsson A. Prediction of transmembrane alpha-helices in prokaryotic membrane proteins: the dense alignment surface method. Protein Eng. 1997; 10: 673-676.

[10] Persson B, Argos P. Prediction of transmembrane segments in proteins utilizing multiple sequence alignments. J Mol Biol. 1994; 237: 182-192.

[11] Rost B, Casadio R, Fariselli P. Topology prediction for helical transmembrane segments at 86% accuracy. Protein Sci. 1996; 5: 1704-1718.

[12] Rost B, Casadio R, Fariselli P, Sander C. Prediction of helical transmembrane proteins at 95% accuracy. Protein Sci. 1995; 4: 521-533.

[13] Sonnhammer EL, von Heijne G, Krogh A. A hidden Markov model for predicting transmembrane helices in protein sequences. Proc Int Conf Intell Syst Mol Biol. 1998; 6: 175-182.

[14] Tusnady GE, Simon I. Principles governing amino acid composition of integral membrane proteins: application to topology prediction. J Mol Biol. 1998; 283: 489-506.

[15] Tusnady GE, Simon I. Topology of membrane proteins. J Chem Inf Comput Sci. 2001; 41: 364-368.

[15] Nugent T, Jones DT. Transmembrane protein topology prediction using support vector machines. BMC Bioinformatics. 2009; 10: 159.

[16] Nugent T, Jones DT. Detecting pore-lining regions in transmembrane protein sequences. BMC Bioinformatics. 2012; 13: 169.

[17] Altaiski M, Mornev M, Polozov R. Wavelet analysis of DNA sequence. Genet Anal. 1996; 12: 165-168.

[18] Hirakawa H, Muta S, Kuhara S. The hydrophobic cores of proteins predicted by wavelet analysis. Bioinformatics. 1999; 15: 141-148.

[19] Li P, Vannucci M. Wavelet change-point prediction of transmembrane proteins. Bioinformatics. 2000; 16: 376-382.

[20] Eisenberg D, Mclachlan AD. Solvation energy in protein folding and binding. Nature. 1986; 319: 199-203.

[21] Huang DS, Xing-Ming Zhao XM, Huang GB,Cheung YM. Classifying protein sequences using hydropathy blocks. Pattern Recognition. 2006; 39: 2293-2300.

[22] Ikeda M, Arai1 M, Okuno T, Shimizu T. TMPDB: a database of experimentally- characterized transmembrane topologies. Nucleic Acids Research. 2003; 31: 406-409.