World Academy of Science, Engineering and Technology
International Journal of Physical and Mathematical Sciences
Vol:4, No:1, 2010

On the Prediction of Transmembrane Helical Segments in Membrane Proteins

Yu Bin, Zhang Yan

***Abstract*—**The prediction of transmembrane helical segments (TMHs) in membrane proteins is an important field in the bioinformatics research. In this paper, a method based on discrete wavelet transform (DWT) has been developed to predict the number and location of TMHs in membrane proteins. PDB coded as 1F88 was chosen as an example to describe the prediction of the number and location of TMHs in membrane proteins by using this method. One group of test data sets that contain total 19 protein sequences was utilized to access the effect of this method. Compared with the prediction results of DAS, PRED-TMR2, SOSUI, HMMTOP2.0 and TMHMM2.0, the obtained results indicate that the presented method has higher prediction accuracy.

***Keywords*—**hydrophobicity, membrane protein, transmembrane helical segments, wavelet transform

I. Introduction

ABOUT 20-30% of genome products have been predicted as membrane proteins, which have significant biological functions in the life activity of the cells [1]. With the proceeding of functional genomics and proteomics research, increasing TM protein sequences are ready to be analyzed, meanwhile, the efficient and high accuracy algorithms are urgently needed to predict TMHs and orientation of transmembrane helices. In addition this prediction supplies references to the research of TM proteins. As a result, the structural prediction of TM proteins, especially in case of the TMHs prediction, is arising much interest of scholars all over the world.

So far many transmembrane helical segments (TMHs) predicting algorithms for membrane proteins have been proposed. Kyte and Doolittle proposed a hydrophobicity scale, based on free energy of transfer of each amino acid between organic solvent and water, and introduced a method for the analysis of the hydrophobicity profile that uses a sliding window of 19~20 residues to enable the detection of potential TMHs as peaks in a two dimensional plot [2]. Von Heijne described a conserved region of positively charged amino acids found on the cytoplasmic side of transmembrane segments [3]. Coined "the positive inside rule", this provide the basis for a new predictive method called SOSUI [4] and PRED-TMR [5]. This method integrated hydrophobicity analysis with information assessment of the positive inside rule to locate putative transmembrane segments and assign a topology to these segments. In recent years ,some statistical prediction methods have been developed that including DAS [6], MEMSAT [7], TMAP [8], PHDhtm [9], TMHMM [1] and HMMTOP [10] .Wavelet transform was first introduced into bioinformatics research in 1996 and raised extensive attention immediately [11-14]. For example, discrete wavelet transform has been applied on hydrophobicity signals in order to predict hydrophobic cores in proteins [15]. In this paper, we make full use of the hydrophobicity of amino acids and multiresolution feature of 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. We selected transmembrane protein sequences from F S Cordes et al [16], and which were constructed into independent test sets to predict TMHs. We compared with main prediction results of DAS, PRED-TMR2, SOSUI, HMMTOP2.0, TMHMM2.0, the obtained results indicate that the proposed method has higher prediction accuracy.

YU Bin is with the College of Mathematics and Physics, Qingdao University of Science and Technology, Qingdao, 266061, China (phone: 86-0532-88956917; e-mail: yubin@qust.edu.cn).

ZHANG Yan is with Qingdao University of Science and Technology, Qingdao, 266061, China (e-mail: zy@qust.edu.cn).

II. Materials and Methods

A. Materials

The code membrane proteins of transmembrane data sets founded by F S Cordes et al was chosen [16], which collects a set of membrane protein structure data identified by crystallography or other experimental technologies such that they can be treated as reliable samples. One group of test data sets that contain total 19 protein sequences including 120 TMHS and 3026 amino acid residues. The data can be obtained from http://www.rcsb.org/pdb/.

B. Methods

Proteins are biomacromolecules that are consisted of twenty different amino acids joined with peptide bonds. Different amino acids have different side-chains that define diverse physico-chemical characteristics of different types of amino acids. Hydrophobic effects are of the most importance among the features because the hydrophobic effects determine to a great degree the stability of protein structures [17]. So considering the critical importance of hydrophobicity in holding the secondary and tertiary structures of proteins, we should map the amino acid sequence of protein onto a sequence of hydrophobicity values that are regarded as raw signals for the wavelet analysis. The hydrophobicity values of 20 amino acids are given in Table I.

TABLE I
HYDROPHOBICITY VALUES OF 20 AMINO ACIDS [2]

Amino Acids	A	C	D	E	F	G	H	I	K	L
H-Values	0.62	0.29	-1.05	-0.87	1.19	0.48	-0.40	1.38	-1.35	1.06
Amino Acids	M	N	P	Q	R	S	T	V	W	Y
H-Values	0.64	-0.85	0.12	-0.78	-1.37	-0.18	-0.05	1.08	0.81	0.26

The wavelet transform (WT) is relatively analysis methods with the changeable time-frequency window, which has very good localization properties in the time and frequency intra-areas. Mallat brought out the most important concept multiresolution analysis (MRA) in a discrete wavelet theory as well as fast algorithm of orthonormal wavelet transformMallat algorithm.

Assume that the shifted scaling function $\{\varphi(t-k), k \in Z\}$ and the shifted wavelet functions $\{\psi(t-k), k \in Z\}$ are orthonormal, respectively. Let $\{c_l^0\}$ denote a sequence of hydrophobicity values, and we define a linear combination $f(t)$ of the sequence with scaling functions $\{\varphi(t-k), k \in Z\}$:

$$f(t) = \sum_{k \in Z} c_k^0 \varphi(t-k) \quad (1)$$

According to a wavelet theory, we have another expansion of $f(t)$:

$$f(t) = \frac{1}{\sqrt{2}}\left(\sum_{k \in Z} c_k^1 \varphi(2^{-1}t-k) + \sum_{k \in Z} d_k^1 \psi(2^{-1}t-k)\right) \quad (2)$$

From (1) and (2) and using orthonormality of the scaling and wavelet functions, we can decompose the sequence $\{c_l^0\}$ into low frequency and high frequency components.

$$c_k^1 = \sum_{l \in Z} c_l^0 \overline{h}_{l-2k} \quad (3)$$

and

$$d_k^1 = \sum_{l \in Z} c_l^0 \overline{g}_{l-2k} \quad (4)$$

Repeatedly application of this decomposition, we can deduce

$$c_k^{j+1} = \sum_{l \in Z} c_l^j \overline{h}_{l-2k}, \quad j = 0,1,2,\cdots, \quad (5)$$

and

$$d_k^{j+1} = \sum_{l \in Z} c_l^j \overline{g}_{l-2k}, \quad j = 0,1,2,\cdots, \quad (6)$$

Conversely, we can derive a reconstruction formula form (1) and (2):

$$c_k^j = \sum_{l \in Z} c_l^{j+1} h_{k-2l} + \sum_{l \in Z} d_l^{j+1} g_{k-2l}, \quad j = 0,1,2,\cdots, \quad (7)$$

Above-mentioned formulas can refer to the literature of Mallat [18].

In (5) and (6), the sequences $\{c_k^{j+1}\}$ and $\{d_k^{j+1}\}$ mean low and high frequencies. In this paper, only the first formula (7) is used because as far as most of the protein hydrophobicity signals are concerned, low frequency domain is especially important and it can reflect the general characteristics of signals. However the high frequency domain is always connected with noise and disturbance, so the basic features of signals will be reserved when the high frequency domain is discarded by putting $d_k^{j+1} = 0$. Using (7), we reconstruct a new sequence $\{\tilde{c}_k^j\}$ only from $\{c_k^{j+1}\}$, that is, we utilize low-pass filtering of wavelet transform, study the general trend and set an optimal threshold to locate TMHs. The threshold here is determined by the biggest average prediction accuracy among a set of protein sequences.

In this paper, we adopted the important Daubechies (dbN) wavelet series as mother wavelet and selected db10 as the optimum wavelet base after analyzing the all data of the test dataset as well as reconstruct wavelet from five different scale levels. To reach a high accuracy in the detection of TMHs, our method is dependent upon the post-treatment of the signals obtained after wavelet reconstruction. For convenience, our prediction method is called WavePrd that is coded in MATLAB programming language.

In order to test the accuracy of prediction methods, we study TM proteins from two aspects — TMHs and amino acid residues.

There are three important evaluation indexes: (1) FP (false-positive): the number of wrongly predicted TMHs; (2) FN (false-negative): the number of not-predicted TMHs; (3) Prediction accuracy of TMHs [12]: $Q_p = \sqrt{M * C} \times 100\%$, here 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. The prediction accuracy of TM protein sequences is computed by $Q_t = (N_{TT}/N_{TOT}) \times 100\%$, where N_{TT} is the number of correctly predicted TM protein sequences and N_{TOT} is the number of TM protein sequences in the test sets.

Prediction accuracy of residues is another evaluation index. The calculation fomula is FAAcor=(NAAcor/NAAall)100%, where NAAcor is the number of correctly predicted TMHs residues and NAAall is the total residues.

III. RESULTS AND DISCUSSION

We pick PDB ID 1F88 from test database as an example to describe the prediction process and predict the number and location of TMHs in membrane proteins [19]. This protein sequence has 348 amino acid residues. The original hydrophobicity plots and and the reconstructed wavelet graphs at each scale level are shown in Figure1. As known to all the peaks of wavelet filtering are corresponding to the real TMHs, a

series of predicted TMHs can be obtained with our method.

From Figure 1 it can be seen that the filtering effects are not distinct at the scale level 1, 2 and 3 while the hydrophobicity signals are over-filtered at scale level 5 such that more usable classification information is concealed. However at the scale level 4 the hydrophobic waveform of 1F88 is corresponding well with the real TMHs. The TMHs prediction accuracy

TABLE II
PREDICTED AND OBSERVED TMHS FOR THE 1F88 PROTEIN SEQUENCE

Observed results	Predicted results	Observed results	Predicted results
35-64	35-57	200-225	205-227
71-100	79-100	247-277	254-280
107-139	112-144	286-306	288-304
151-173	153-175		

reaches 100% and the amino acid residues prediction accuracy reaches 95.3% at the scale level 4 with optimal threshold 0.423. The contrast data in Table II show above result more clearly.

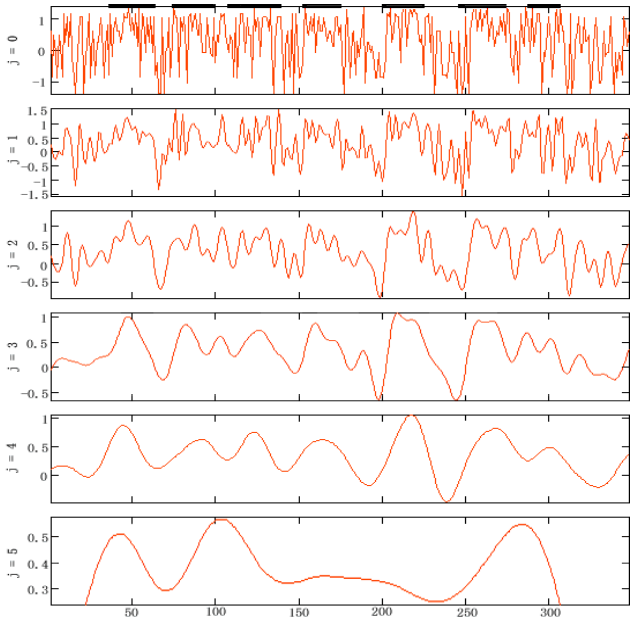

Fig. 1 The hydrophobicity signal plot and low frequencies at five different scale levels for 1F88 protein

After the analysis of 19 set of membrane protein, db10 is selected as the optimal wavelet. The optimal threshold is 0.423 at the scale level 4, 115 of 121predicted TMHs are the true TMHs. So we can gain results as following: the average TMHs prediction accuracy is 95.4%, the average prediction accuracy for residues is 74.8%, the number of false-positive TMHs is 6 and the number of false-negative TMHs is 5. We predict 19 set of membrane proteins by 5 methods—DAS [6], HMMTOP2.0 [10], PRED-TMR2 [5], SOSUI [4], TMHMM2.0 [1] and the prediction result can be found in Table3.

From Table III, HMM-based TMHMM2.0 has the highest TMHs prediction accuracy that reaches 95.8%, the next two are WavePrd and HMMTOP2.0 with prediction accuracy 95.4% and 94.1%, and the lowest prediction accuracy of statistical methods DAS is 87.0%, which is 8.4% lower than our method. Amino acid residues prediction accuracy of TMHMM2.0 is highest, which reaches 77.9%, that of HMMTOP2.0 is 74.6.% and WavePred is 74.8%. Compared with several prediction methods, our method is more accurate and effective in predicting the TMHs number and location of membrane proteins.

IV. CONCLUSION

The study of the structure and function of TM proteins is increasingly emphasized since TM proteins play an extraordinarily important role in the life activity of the cells, such as signal transduction, immune response and membrane transport. The computer prediction and analysis of the TMHs is able to provide much important information to disclose the relationship between the structure and function of TM proteins. In this paper, a method based on discrete wavelet transform (DWT) has been developed to predict the number and location of TMHs in membrane proteins. Compared with the most prediction results of several prediction methods, the obtained results indicate that the presented method has higher prediction accuracy.

ACKNOWLEDGMENT

This work was supported by grants from the National Natural Science Foundation of China (No. 3057l059).

TABLE III
MAIN RESULTS OF SIX PREDICTION METHODS

Method	Nobs	Nprd	Ncor	Q_p	M	C	Ntot	Ntt	Q_t	FP	FN	FAcor
WavePrd	120	121	115	95.4%	95.8%	95.0%	19	13	68.4%	6	5	74.8%
DAS	120	133	110	87.0%	91.7%	82.7%	19	10	52.6%	23	10	62.3%
HMMTOP2.0	120	118	112	94.1%	93.3%	94.9%	19	14	73.7%	6	8	74.6%
PRED-TMR2	120	101	96	87.2%	80.0%	95.1%	19	8	42.1%	5	24	61.6%
SOSUI	120	111	107	92.7%	89.2%	96.4%	19	12	63.2%	4	13	74.1%
TMHMM2.0	120	116	113	95.8%	94.2%	97.4%	19	13	68.4%	3	7	77.9%

REFERENCES

[1] A. Krogh, B. Larsson, G. von Heijne, E. Sonnhamme, "Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes,"*J. Mol. Biol.*, vol. 305, 2001, pp. 567-580.
[2] J. Kyte, R. F. Doolittle, "A simple method for displaying the hydrophathic character of a protein," *J. Mol. Biol.*, 1982, 157: 105-132.
[3] G. Heijne, "The distribution of positively charged residues in bacterial inner membrane proteins correlates with the transmembrane topology," *EMBO J*, vol. 5, 1986, pp. 3021-3027.
[4] T. Hirokawa, S. Boon-Chieng, S. Mitaku, "SOSUI: classification and secondary structure prediction system for membrane proteins," *Bioinformatics*, vol. 14, 1998, pp. 378-379.
[5] C. Pasquier, V. J. Promponas, G. A. Palaios, J. S. Hamodrakas, S. J. Hamodrakas, "A novel method for predicting trsnsmembrane segments in proteins based on a statistical analysis of the SwissProt database: the PRED-TMR algorithm," *Protein Eng*., vol. 12, 1999, pp. 381-385.
[6] M. Cserzö, E. Wallin, I. Simon, G. von Heijne, A. Elofsson, "Prediction of transmembrane alpha-helices in prokaryotic membrane proteins: the dense alignment surface method," *Protein Eng*., vol. 10, 1997, pp. 673-676.
[7] D. T. Jones, W. R. Taylor, J. M. Thornton, "A model recognition approach to the prediction of all-helical membrane protein structure and topology," *Biochemistry*, vol. 33, 1994, pp. 3038-3049.
[8] B. Persson, P. Argos, "Prediction of transmembrane segments in proteins utilizing multiple sequence alignments," *J. Mol. Biol.*, vol. 237, 1994, pp. 182-192.
[9] B. Rost, R. Casadio, P. Fariselli, "Topology prediction for helical transmembrane segments at 86% accuracy," *Protein Sci*., vol. 5, 1996, pp. 1704-1718.
[10] G. E. Tusnady, I. Simon, "Principles governing amino acid composition of integral membrane proteins: application to topology prediction," *J. Mol. Biol*., vol. 283, 1998, pp. 489-506.
[11] Altaiski, M. Mornev, O. Polozov, "Wavelet analysis of DNA sequence," *Genet. Anal*. , vol. 12, 1996, pp. 165-168.
[12] B. Yu, X. H. Meng, H. J. Liu, et al, "Prediction of transmembrane helical segments in transmembrane proteins based on wavelet transform," *Journal of Shanghai University (English Edition)*, vol. 10, 2006, pp. 308-318.
[13] P. Liò, "Wavelets in bioinformatics and computational biology: state of art and perspectives," *Bioinformatics*, vol. 19(1) 2003, pp. 2-9.
[14] J. P. Mena-Chalco, Y. Zana, and R. M. Cesar, "Identification of protein coding regions using the modified Gabor-wavelet transform," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 5, 2008, pp. 198-207.
[15] H. Hirakawa, S. Muta, S. Kuhara, "The hydrophobic cores of proteins predicted by wavelet analysis," *Bioinformatics*, vol. 15, 1999, pp. 141-148.
[16] F. S. Cordes, J. N. Bright, M. S. Sansom, "Proline-induced distortions of transmembrane helices," *J. Mol. Biol*., vol. 323, 2002, pp. 951-960.
[17] D. Eisenberg, A. D. Mclachlan, "Solvation energy in protein folding and binding," *Nature*, vol. 319, 1986, pp. 199-203.
[18] S. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE Trans. Pattern Anal. Math.Intell*, vol. 11, 1989, pp. 674-693.
[19] K. Palczewski, T. Kumasaka, T. Hori, "Crystal structure of rhodopsin: A G protein-coupled receptor," *Science*, vol. 289, 2000, pp. 739-745.

YU Bin received the ME degree in computational mathematics from Shanghai University, Shanghai, China, in 2005. His research interests include bioinformatics, systems biology and matrix theory.

ZHANG Yan received the BE degree in computer science and technology from Northwestern Polytechnic University, Xi'an Shanxi, China, in 2004, the ME degree in control theory and control engineering from the Qingdao University of Science and Technology, Qingdao, Shandong, China, in 2009. His research interests include wavelet theory, pattern recognition and image processing.