Chinese Science Bulletin

A new approach to the prediction of transmembrane structures

LIU HongDe[1], WANG Rui[1], LU XiaoQuan[1†], CHEN Jing [1], LIU XiuHui[1] & DING Lan[2]

[1] College of Chemistry and Chemical Engineering, Northwest Normal University, Lanzhou 730070, China;
[2] College of Life Science, Northwest Normal University, Lanzhou 730070, China

About 20%—30% of genome products have been predicted as membrane proteins, which have significant biological functions. The prediction of the amount and position for the transmembrane protein helical segments (TMHs) is the hot spot in bioinformatics. In this paper, a new approach, maximum spectrum of continuous wavelet transform (MSCWT), is proposed to predict TMHs. The predictions for eight SARS-CoV membrane proteins indicate that MSCWT has the same capacity with software TMpred. Moreover, the test on a dataset of 131 structure-known proteins with 548 TMHs shows that the prediction accuracy of MSCWT for TMHs is 91.6% and that for membrane protein is 89.3%.

maximum spectrum of continuous wavelet transform (MSCWT), membrane protein, transmembrane protein helical segments (TMHs)

As the important components of nerve signal molecule, hormone, acceptors and all kinds of ion channels, membrane proteins play an important role in the life activity of the cells[1]. They are also the binding sites of some drugs[2]. However, directly measuring 3-D structure of the membrane proteins with experimental way (X-ray crystal diffraction) is not easy because their stable natural conformations often need the assistant of the biology membrane. Therefore, it is one of the important bioinformatics subjects to develop a highly effective and accurate approach to predicting the structure of the membrane protein.

In literature[3], protein sequence was converted into a digital signal by substituting amino acid residues with their hydrophobic free energy. Then, a sliding window was used to scan the signal. The probable TMHs were predicted by a reasonable threshold. Von Heijine proposed the well-known 'positive-inside rule'[4], which provided a further guide for prediction. Multiple sequence alignment information of membrane proteins was also used in such prediction[5]. From view of method used, besides of neural networks and sliding window technique[4,5], Markov model was also used[6]. In a Markov model, seven states were designed to respond to seven regions of membrane protein, namely, mohelix core, cap cyt, cap non-cyt, loop cyt, short loop non-cyt, long loop non-cyt and globular[6]. In literature[7], the inclinations that each amino acid occurred in seven regions were computed. Then, the prediction was achieved by dynamic programming algorithm. Similarly with literature[7], literature[8,9] also achieved the prediction based on amino acid occurrence frequency. The improvement was the prediction for transmembrane directions.

Wavelet transform is called "the mathematical microscope"[10,11], which has the good capacity in dealing with non-stationary signal and singular signals. Its first application in bioinformatics was the prediction for the protein hydrophobic core[12]. It was also found that wavelet transform has the capacity of recognizing the different structures of proteins[13]. The detection of protein motifs can also be achieved by wavelet transform[14].

Received March 31, 2007; accepted October 24, 2007
doi: 10.1007/s11434-008-0055-5
†Corresponding author (email: luxq@163.com)
Supported by the National Natural Science Foundation of China (Grant Nos. 20775060 and 20335030), the Teaching and Research Award Program for Outstanding Young Teachers in Higher Education Institutions of China, and the Key Lab of Polymer Material Science of Gansu Province, China

Recently, a wavelet-based transmembrane prediction was reported, in which the prediction was based on the single-scale continuous wavelet transform (CWT)[15].

In this paper, MSCWT was proposed to predict TMHs of membrane proteins. The method has been successfully applied in resolving overlapped signals including high performance liquid chromatography signals, differential pulse voltammetric signals, ultraviolet-visible spectrum signals and so on[16]. Differing from single-scale maximum detection, MSCWT is based on multi-scale CWT. It has the capacity of detecting the maximum on different scales simultaneously, which is very helpful in analyzing multi-frequency constituent signals.

1 Data and method

1.1 Data

The test dataset is retrieved from the latest MPtopo database (http://blanco.biomol.uci.edu/mptopo/). The protein structures have been resolved by crystallography. Eight SARS-CoV membrane protein sequences are retrieved form the website http://athena.bioc.uvic.ca/sars/map/diagram.html. They are orf3, orf4, orf7, orf8, orf9, orf14, m and s.

1.2 Method

Given a signal *f(t)* of time domain, its CWT is expressed with eq. (1),

$$Wf(a,b)=\frac{1}{\sqrt{a}}\int_{-\infty}^{+\infty} f(t)\psi_{a,b}^{*}(t)\mathrm{d}t\,, \qquad (1)$$

where Ψ is the wavelet, a and b is the scale and the translation, respectively. $Wf(a,b)$ is a function of a and b.

MSCWT is derived from CWT. It is designed to locate signals' maximum positions, especially for multi-frequency constituent signals. MSCWT of a signal could be obtained from the following steps: firstly, choose an appreciated mother wavelet and an appreciated scale range to perform CWT; then, detect and record the CWT maximum at every translation; finally, plot the recorded maximum value to its position (translation). The following is the MSCWT pseudocode:

****

For b =1 to length of signal

  For a =a1 to a2 'a1 and a2 form the range of scale'

  Find and record the maximum of Wf(a, b)

  Next a

  Plot the maximum to b

Next b

Notes: a1 and a2 refer to the bottom and top limit of the scale range, respectively; $Wf(a,b)$ are CWT coefficients.

****

In this paper, morlet function is chosen as the analytical wavelet. To get MSCWT with a higher resolution, scales are set from 1 to 64.

The amounts and positions of TMHs are obtained as the following steps:

Step 1: Convert protein sequence into a raw signal by substituting amino acid residue with its hydrophobic free energy[3].

Step 2: Perform CWT for the raw signal and obtain MSCWT.

Step 3: Set 1.5 as threshold and find regions more than the threshold in MSCWT as the candidates of TMHs. If the length of the candidate is less than 20 amino acid residues, the candidate is not TMHs because it is too short. If two or more TMHs are very close (less than three amino acids), they are combined a whole TMHs.

The prediction is evaluated at three aspects: amino acid, TMHs and protein sequence[17]:

(1) The prediction accuracy of amino acid residues

The prediction accuracy rate of amino acid residues is calculated by $F_{\mathrm{AAcor}}=(N_{\mathrm{AAcor}}/N_{\mathrm{AAall}})\times 100\%$ where N_{AAcor} is the number of correctly predicted amino acid residues; N_{AAall} is the total number of amino acid residues predicted.

(2) The prediction accuracy of TMHs

i. *FP* (false-positive): the number of wrongly predicted TMHs;

ii. *FN* (false-negative): the number of not predicted TMHs;

iii. Prediction accuracy of TMHs: $Q_P=\sqrt{M\times C}\times 100\%$, where $M=N_{\mathrm{cor}}/N_{\mathrm{obs}}$ (N_{cor} is the number of correctly predicted TMHs, N_{obs} is the number of observed TMHs), and M can be regard as a measure index of sensitivity; $C=N_{\mathrm{cor}}/N_{\mathrm{prd}}$ (N_{prd} is the total number of predicted TMHs), and C is regarded as a measure index of specificity.

(3) The prediction accuracy of membrane protein sequences

Once all the transmembrane regions of a membrane protein sequence are predicted properly, the whole membrane protein sequence is considered to be predicted correctly. The prediction accuracy of membrane

protein sequences is: $Q_t = (N_{TT}/N_{TOR}) \times 100\%$, where N_{TT} is the number of correctly predicted membrane protein sequences and N_{TOR} is the number of membrane protein sequences in the test sets.

The predictions by MSCWT are compared with those by other two software, Tmpred (http://www.ch.embnet.org/software/TMPRED_form.html) and DAS (http://www.sbc.su.se/~miklos/DAS/).

2 Results and discussion

MSCWT can be used to predict the existence of TMHs and provide the information about the beginning position and end position. Moreover, it can be used to analyze the finer structures of proteins.

Figure 1 shows the TMHs predicted by MSCWT and TMpred for SARS-CoV protein Orf9 and Orf14. For the predictions of TMpred (subplot a' and b'), solid line and dotted line represent two model signals. i→o and o→i indicate the direction from inner membrane to outer membrane and vice versa, respectively. The predictions for eight SARS-CoV proteins are listed in Table 1. From Figure 1 and Table 1, it is concluded as follows: (1) MSCWT amplifies the profile of hydrophobic and hydrophilic residues, which is helpful in observing the changes of protein hydrophobia. The great peaks in MSCWT respond to the changes of hydrophobias, which is the characteristics of TMHs. In fact, the MSCWT curve more than the threshold (1.5) denotes TMHs. Figure 1 shows that the predictions by MSCWT are in consistence with those by TMpred, indicating MSCWT has a good precision. (2) In this paper, CWT is performed on a scale range, which makes that the large scale corresponds to lower frequencies of signals while small scale to high frequencies. Thus, MSCWT can be used to study the global properties of signals as well as local properties. In MSCWT, large span peaks are not very smooth and contain some small peaks, indicating the hydrophobias vary with residue positions in TMHs, which suggests that the protein has fine folding structure in TMHs and such information can be observed by MSCWT. This is why those small peaks are not filtered. Our experi-

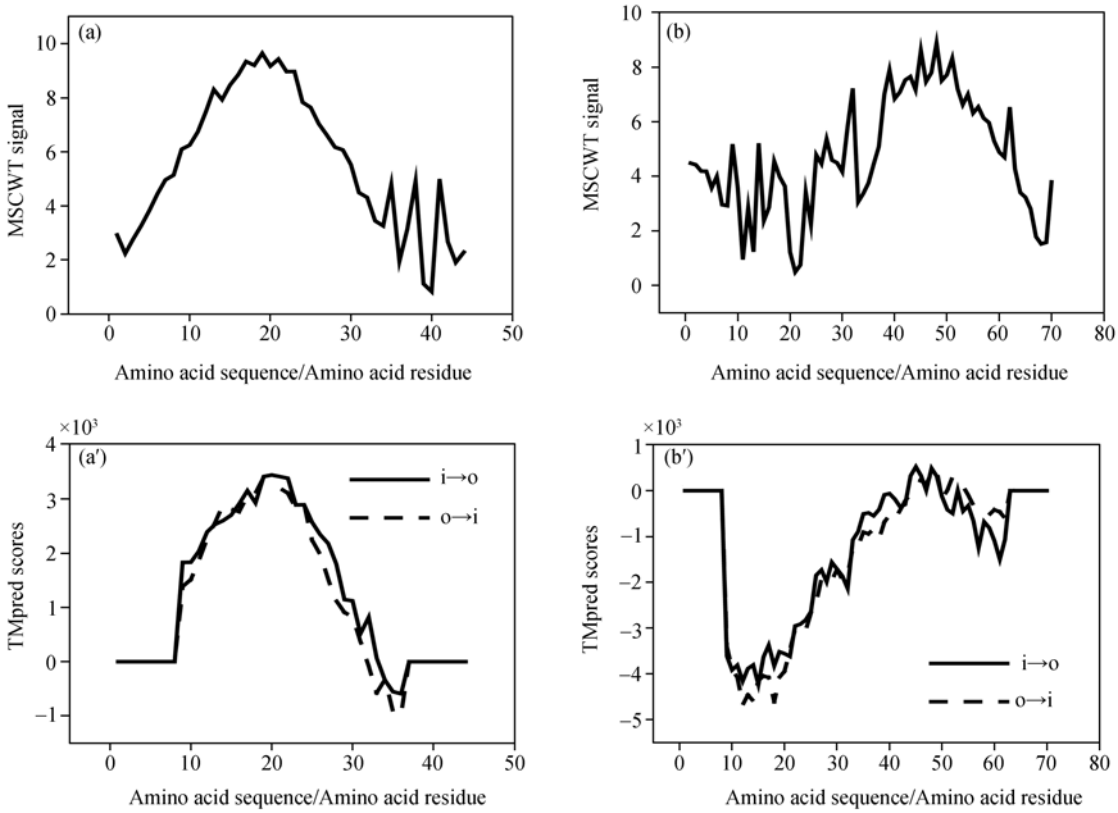

Figure 1 Comparison of MSCWT and TMpred in predicting the SARS-CoV proteins Orf9 and Orf14. The longitudinal axis is defined by forecast signals, and the transverse axis is defined by points at transmembrane protein sequences. (a) MSCWT for Orf9, TMHs predicted is from 8 to 30; (b) MSCWT for Orf14, TMHs predicted is from 36 to 58; (c) Tmpred for Orf9, TMHs predicted is from 9 to 30; (d) Tmpred for Orf14, TMHs predicted is from 36 to 58. In TMpred, solid line and dotted line represent two signals suggested, i→o indicates the direction from inner membrane to outer membrane, and o→i indicates the reverse direction.

Table 1 Comparison of MSCWT and TMpred in predicting TMHs of SARS-CoV proteins, the threshold of MSCWT is 1.5

Protein	Method	TMHs 1	TMHs 2	TMHs 3	TMHs 4
Orf7	MSCWT	7-29	32-41		
	TMpred	7-29	21-41		
Orf8	MSCWT	1-14	47-68		99-117
	TMpred	1-18			99-117
Orf9	MSCWT	8-30			
	TMpred	9-30			
Orf4	MSCWT	4-20			52-75
	TMpred	1-20			52-75
Orf14	MSCWT	36-58			
	TMpred	36-58			
Orf3	MSCWT	37-54		92-121	
	TMpred	41-58	76-99	86-121	
s	MSCWT	1-15	55-85	233-240:575-594	309-325:530-540
	TMpred	1-17	51-73	233-257:570-88	345-364:524-542
m	MSCWT	21-32	47-63	72-95	125-146
	TMpred	21-38	51-69	74-96	

ment shows that if the bottom scale is set to 8, a smooth and global hydrophobias profile will be observed.

In order to test the reliability of MSCWT, a dataset of 131 structure-known membrane proteins containing 548 TMHs is constructed based on MPtopo data bank. MSCWT reports 530 transmembrane regions, in which, 81 regions are false-positive. 18 regions are not predicted yet, and the leakage rate is 3.3%. Among 131 protein sequences, 117 sequences are correctly predicted. The accuracy rate of TMHs is 91.6% and the accuracy rate for membrane protein sequences is 89.3%.

In addition, 65 membrane proteins are chosen from the above test dataset as the new testing dataset in their original order. The new dataset is predicted by software TMpred and DAS, and the results are shown in Table 2. The difference among these methods is not obvious. Compared with the two methods, the leakage amount of MSCWT is 13. The number of TMpred and DAS are 33 and 30, respectively.

Table 2 Prediction performance for MSCWT, TMpred and DAS

Method	N_{obs}	N_{prd}	N_{cor}	FP	FN	Q_p	N_{TOR}	N_{TT}	Q_t
MSCWT	277	313	264	49	13	89.7%	65	55	84.6%
TMpred	277	248	244	4	33	93.1%	65	49	75.4%
DAS	277	278	247	31	30	89.0%	65	52	80.0%

For the accuracy rate of TMHs, TMpred is 93.1%; MSCWT is 89.7%; and DAS is 89.0%. MSCWT has the highest accuracy rate of membrane protein sequence (84.6%), while TMpred and DAS are 75.4% and 80.0%, respectively. All the above indicates that MSCWT has a relatively high performance among the existing methods.

3 Conclusions

In this paper, the proposed new method MSCWT shows a good efficiency in predicting the positions and amounts of TMHs of membrane protein. The limit of MSCWT is that it cannot provide information about the direction of TMHs.

1 Tanford C. How protein chemists learned about the hydrophobic factor. Protein Sci, 1997, 6(6): 1358－1366
2 Adams P D, Arkin I T, Engelman D M, et al. Computational searching and mutagenesis suggest a structure for the pentameric transmembrane domain of phospholamban. Nat Struct Biol, 1995, 2(2): 154－162
3 Kyte J, Doolittle R F. A simple method for displaying the hydropathic character of a protein. J Mol Biol, 1982, 157(1): 105－132
4 von Heijne G. Membrane protein structure prediction. Hydrophobicity analysis and the positive inside rule. J Mol Biol, 1992, 225(2): 487－494
5 Rost B, Casadio R, Fariselli P. Transmembrane helices predicted at 95% accuracy. Protein Sci, 1995, 4(3): 521－533
6 Sonnhammer E L, 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
7 Jones D T, Taylor W R, Thornton J M. A model recognition approach to the prediction of all-helical membrane protein structure and topology. Biochemistry, 1994, 33(10): 3038－3049
8 Persson B, Argos P. Prediction of transmembrane segments in proteins utilizing multiple sequence alignments. J Mol Biol, 1994, 273(2): 182－192
9 Persson B, Argos P. Topology prediction of membrane proteins. Protein Sci, 1996, 5(2): 363－371
10 Lu X Q, Mo J H. Wavelet analysis as a new method in analytical chemometrics. Chin J Anal Chem, 1996, 24(9): 1100－1106
11 Lu X Q, Liu H D. Wavelet Analysis Techniques in Analytical Chemestry. Beijing: Chemistry Industry Press, 2006
12 Hirakawa H, Muta S, Kuhara S. The hydrophobic cores of proteins predicted by wavelet analysis. Bioinformatics, 1999, 15(2): 141－148
13 Mandell A J, Selz K A, Shlesinger M F. Wavelet transformation of protein hydrophobicity sequences suggests their memberships in structural families. Physica A, 1997, 244: 254－262
14 Murray K, Gorse D, Thornton J. Wavelet transforms for the characterization and detection of repeating motifs. Mol Biol, 2002, 316(2): 341－363
15 Qiu J D, Liang R P, Mo J Y, et al. Prediction of transmembrane proteins based on the continuous wavelet transform. Chem Inf Comput Sci, 2004, 44(2): 741－747
16 Lu X Q, Liu H D, Xue Z H, et al. Maximum spectrum of continuous wavelet transform and its application in resolving an overlapped signal. J Chem Inf Comput Sci, 2004, 44(4): 1228－1237
17 Chen Z Q , Liu Q, Zhu Y S, et al. Performance analysis of methods that predict transmembrane regions. Acta Biochim Biophys Sin, 2002, 34(3): 285－290