Available online www.jocpr.com

Journal of Chemical and Pharmaceutical Research, 2013, 5(9):465-471

Research Article

ISSN : 0975-7384
CODEN(USA) : JCPRC5

Research on prediction of transmembrane protein topology based on fuzzy theory

Xian Wen Luo[1] Zuo Ying Liu[2]

[1]*Information Management Department Southwest University Chongqing, China*
[2]*Department of Basic Sciences, Southwest University Chongqing, China*

ABSTRACT

There are many methods proposed to predict transmembrane protein topology. However, these methods have some shortcomings to correctly predict the boundary of the transmembrane region. In this paper, a new transmembrance prediction based on data fusion technology is proposed. The results of different prediction methods, described as different interval number, can be combined with induced ordered weighted averaging operator to determine the appropriate fuzzy interval. The appropriate fuzzy interval can be determined. The transmembrane regions can be predicted by choosing the threshold. The effectiveness of the transmembrane prediction is verified by experiments.

Key words: Transmembrane protein topology prediction; Fuzzy set theory; IOWA operator; Information fusion

---

INTRODUCTION

Transmembrane protein is a particular and important class in proteins, the particularity is that it spans the phospholipid bilayer of cell membrane. Transmembrane protein is important because it gives the membrane a variety of functions. Some transmembrane proteins can be used as "carrier" to transport material in or out of cells, some transmembrane proteins are specific receptors for hormones or other chemicals, and some transmembrane proteins determine the identification function of cells. Because of this, the research on the entire membrane protein topology is very meaningful. However, so far only a small part of the transmembrane protein structure is known. Although the multi-dimensional nuclear magnetic resonance (NMR), X-ray crystallography, electron diffraction of two-dimensional and three-dimensional image reconstruction technology, provide an effective test method for the determination of protein structure, these methods are time consuming and require quantity and purity of the sample so their applications subject to considerable restrictions. Therefore, the study on simple prediction method of protein structure is very urgent, and it is an important research field in bioinformatics.

So far, there are two known structures of transmembrane protein: the helix bundle formed by α–helices and the barrel structure folded by β–barrel. This paper focuses on the prediction methods of topology structure of α–helical transmembrane protein.

The α–helical membrane protein has two significant features, and the currently existing prediction methods are directly or indirectly developed on the basis of these two features. The two notable features are as follows: [1] The transmembrane region has a strong hydrophobic. [2] "Positive-inside rules", that is, Lys and Arg --- positively charged amino acids are located in the inner (intracellular), while extremely rare in the outer (extracellular).

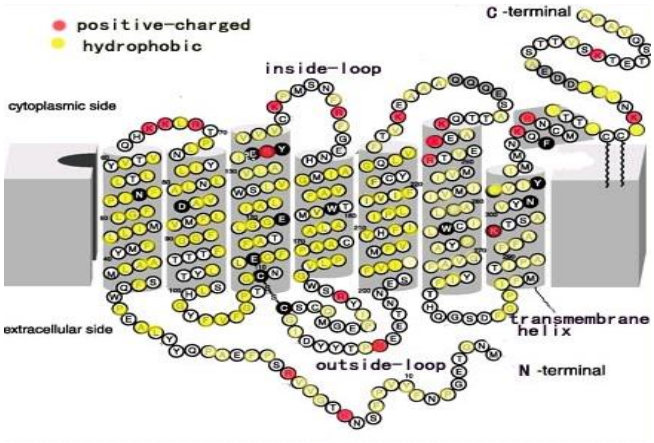

Fig 1.1 Topology model of Rhodopsin. Hydrophobic residues are shown in yellow circles and positive-charged residues are shown in red circles

These two features of the spiral membrane proteins provide a basis for the prediction of the topology structure of transmembrane protein. The first feature is applied to the prediction of transmembrane regions, while the second feature makes it possible to predict the transmembrane direction. The following topology prediction methods are developed on the basis of these two features.

2. Basic theories

2.1 Fuzzy set theory

The concept of fuzzy set is first proposed in 1965 by Chad, a expert in cybernetics in California University. The emergence of the concept “fuzzy sets” makes it possible to use the mathematical thinking and methods to handle fuzzy phenomenon, which constitutes the basis of fuzzy set theory.

Fuzzy sets can be defined as follows: suppose a mapping is given in a domain U,

$$A: U \rightarrow [0,1] \qquad u \mid\rightarrow A(u) \qquad 2\text{-}1$$

then A is the fussy set in domain U, A (u) is called the membership function of A or the membership degree of u to A. For simplicity, "Fuzzy " is recorded as "F", that is, "fuzzy sets", is written as "F set.”.

In objective things, the most common is the case in which a real number R as the domain, the membership function of F set on R is called F distribution. According to the nature of the problem, the appropriate distribution (ie in line with the actual situation) can be selected, then the determination of the membership function is very simple. Starting from the actual situation of the transmembrane protein, apparently the distribution can meet the trapezoidal distribution.

For the trapezoidal fuzzy number $\tilde{A} = (a_1, a_2, a_3, a_4)$ its membership function is defined as

$$\mu_{\tilde{A}} = \begin{cases} 0 & x < a_1 \\ \dfrac{x - a_1}{a_2 - a_1} & a_1 \le x < a_2 \\ 1 & a_2 \le x < a_3 \\ \dfrac{a_4 - x}{a_4 - a_3} & a_3 \le x < a_4 \\ 0 & a_4 \le x \end{cases} \qquad 2\text{-}2$$

For transmembrane proteins, the interval (0, a_1) indicates the amino acid sequence which does not belong to of the transmembrane region, its membership is 0. Interval [a_1, a_2) indicates that the membership degree to which the amino acid sequences of this interval belong to the transmembrane region is gradually increasing. The interval $[a_2, a_3)$ indicates that the amino acid sequence in this region must belong to the transmembrane region, and its membership degree is 1. The interval $[a_3, a_4)$ indicates that the membership degree to which the amino acid sequences of this interval belong to the transmembrane region is gradually decreasing. The area greater than a_4 indicates the amino acid sequence is sure not to belong to of the transmembrane region, its membership is 0.

2.2.Brief introduction to the operator IOWA

Induced ordered weighted average (on IOWA) operator [15] was proposed by Yager, which is the expansion of induced weighted average (OWA) operator [16]. By empowering in order each time point in the sample interval according to the accuracy degree of each prediction method and taking the error sum of squares as a criterion, a new combination forecasting model is established. It is defined as follows:

suppose $< v_1, a_1 >, < v_2, a_2 >, \cdots, < v_m, a_m >$ is m -dimensional array, and

$$f_w(< v_1, a_1 >, < v_2, a_2 >, \cdots < v_m, a_m >) = \sum_{i=1}^{m} \omega_i a_{v-index(i)} \quad 2\text{-}3$$

then f_w is m -dimensional induced ordered weighted averaging operator resulting from $v_1, v_2, \cdots, v_m$, abbreviated as IOWA, v_i is called the induced valueof a_i. $v - index(i)$ is the subscript of listing i large number arranged in descending order , $W = (\omega_1, \omega_2, \cdots, \omega_m)^T$ is the weight vector in OWA , meet. $\sum_{i=1}^{m} \omega_i = 1, \omega_i \geq 0, i = 1, 2, \cdots, m$.

2.3 The data fusion algorithm

The data fusion algorithm uses the information contained in the data itself to avoid the pre-set threshold, thereby reduces the subjectivity of the algorithm and improved its operability. The relative distance between the data is defined as d_{ij} , and its form is expressed as

$$d_{ij} = | x_i - x_j | \quad 2\text{-}4$$

The form shows that the larger d_{ij} is, the smaller the degree of mutual support between the two data is. And then define a support function r_{ij} which satisfies two conditions by itself:

① r_{ij} is inversely proportional to the relative distance;

② $r_{ij} \in (0,1]$ enables data processing to take advantage of the advantages of membership function in fuzzy set theory, to avoid the absoluteness of the mutual support degree between data. Then the support function r_{ij} is defined as

$$r_{ij} = -\frac{d_{ij}}{\max\{d_{ij}\}} + 1 \quad 2\text{-}5$$

For the data fusion problem, establish the support matrix R

$$R = \begin{Bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{nn} \end{Bmatrix} \quad 2\text{-}6$$

If a message can be shared by a number of sub-systems, then $\sum_{i=1}^{n} \varpi_i = 1$

ω_i should integrate the overall information of $r_{i1}, r_{i2}, \cdots, r_{in}$,by the merge theory of the probability of source,

which requires a group of non-negative $v_1, v_2, \cdots, v_n$ to meet the equation

$$\varpi_i = v_1 r_{i1} + v_2 r_{i2} + \cdots + v_n r_{in} \qquad 2\text{-}7$$

Rewrite the form into a matrix, then

$$W = RV \qquad 2\text{-}8$$

In the matrix, $W = [\varpi_1, \varpi_2, \cdots, \varpi_n]^T, V = [v_1, v_2, \cdots, v_n]$, $r_{ij} \geq 0$, so the support matrix is a non-negative matrix. According to the property of non-negative matrix, R has the maximum modulus eigenvalue $\lambda \geq 0$. From $\lambda V = RV$, its corresponding eigenvector can be obtained $V = [v_1, v_2, \cdots, v_n]^T$. Suppose

$$\varpi_i = \frac{v_i}{v_1 + v_2 + \cdots + v_n} \qquad 2\text{-}9$$

Then ϖ_i is the weight coefficients of number i measured datum, and the integrated results of n measured data are

$$x = \varpi_1 x_1 + \varpi_2 x_2 + \cdots + \varpi_n x_n \qquad 2\text{-}10$$

3. New prediction method

3.1. Data sets and basic algorithm

MPtopo is a transmembrane protein database, it stores the amino acid sequence of the transmembrane protein in the known topology structure. In this article, 124 amino acid sequence are used. They were divided into 10 groups and 10-fold cross-validation is carried out, and nine of them as a training set and one of them as a test set in turn. To integrate different algorithms, five kinds of transmembrane protein prediction algorithm will be used as a basis, they are OCTOPUS、PRO-TMHMM、PRODIV-TMHMM、SCAMPI-mas and SCAMPI-seq.

3.2 Model establishment

For the same membrane protein sequences, there exist some differences in prediction results of the five kinds of transmembrane protein prediction algorithm, and these differences emerge in the boundary of amino acid sequence of transmembrane protein. It is necessory to integrate the results of the five kinds of transmembrane protein prediction algorithm. Suppose the value of the amino acid sequence of real transmembrane protein is x_t.then x_t indicates the prediction result of the t time of the prediction algorithm i.

$$k_{it} \begin{cases} 1-|x_t - x_{it}| & |(x_t - x_{it})/x_t| < 1 \\ 0 & |(x_t - x_{it})/x_t| \geq 1 \end{cases} \qquad 3\text{-}1$$
$$i = 1,2,\cdots,5 \qquad t = 1,2,\cdots,N$$

then k_{it} indicates the prediction accuracy of the t time of the prediction algorithm i, apparently $k_{it} \in [0,1]$. Take prediction accuracy k_{it} as the induced value of the predictive value x_{it}, so that the prediction accuracy of the t time of a single prediction method in i kinds and its corresponding predicted values for constitute a five two-dimensional arrays: $< k_{1t}, x_{1t} >, < k_{2t}, x_{2t} >, \cdots, < k_{5t}, x_{5t} >$.

Suppose $W = (\omega_1, \omega_2, \cdots, \omega_5)^T$ is ordered weighted averaging vector of various forecasting methods in combination forecast, list the prediction accuracy sequences $k_{1t}, k_{2t}, \cdots, k_{5t}$ of the t time of the five kinds of

individual prediction method in descending order, set $k-index(it)$ as the subscript of the prediction accuracy of the t time of number i, according to the definition of induced ordered weighted average (on IOWA) operator, set

$$IOWA(<k_{1t},x_{1t}>,<k_{2t},x_{2t}>,\cdots,<k_{5t},x_{5t}>)=\sum_{i=1}^{5}\omega_i x_{k-index(it)} \quad 3\text{-}2$$

, then equation 3-2 is called the predictive value of the induced ordered weighted averaging combination resulting from the prediction accuracy sequence.

Obviously equation 3-2 has nothing to do with individual prediction methods, but with the size of the prediction accuracy of the individual prediction methods at each time point.

Set $e_{k-index(it)}=x_t-x_{k-index(it)}$ then the error sum of squares S of 5 combination of prediction is:

$$S=\sum_{i=1}^{5}(x_t-\sum_{i=1}^{5}\omega_i x_{k-index(it)})^2=\sum_{i=1}^{5}\sum_{j=1}^{5}\omega_i\omega_j(\sum_{i=1}^{5}e_{k-index(it)}e_{k-index(jt)}) \quad 3\text{-}3$$

The new forecasting model of induced ordered weighted average combination can be expressed as the following model:

$$\min S(\omega_1,\omega_2,\cdots,\omega_5)=\sum_{i=1}^{5}\sum_{j=1}^{5}\omega_i\omega_j(\sum_{i=1}^{5}e_{k-index(it)}e_{k-index(jk)}) \quad 3\text{-}4$$

$$s.t.\begin{cases}\sum_{i=1}^{5}\omega_i=1\\ \omega_i\geq 0, i=1,2,\cdots,5\end{cases}$$

MATLAB optimization toolbox can be used to calculate the optimal weight The above formula can be used to obtain the prediction interval [c1,d1],[c2,d2],…[ct,dt] after the new integration of transmembrane protein. Then compare and the actual interval of the transmembrane protein[c1,d1],[c2,d2]…[ct,dt] to get a fuzzy interval number of the fusion. The same method can be used to deal with the right border range. Table 3-1 shows the training results of 124 transmembrane protein.

Table 3-1 Training result

Training set	ω_1	ω_2	ω_3	ω_4	ω_5	Fuzzy interval
First fold	0.3553	0.2964	0.1614	0.1191	0.0678	[-1.6383,3.5594]
Second fold	0.4017	0.3315	0.1087	0.0897	0.0684	[-3.6303,1.7208]
Third fold	0.3925	0.3311	0.1158	0.0963	0.0643	[-3.5933,1.5888]
Fourth fold	0.3792	0.3203	0.1304	0.1036	0.0665	[-3.3487,1.6852]
Fifth fold	0.3933	0.3335	0.1152	0.0952	0.0638	[-3.2849,1.5970]
Sixth fold	0.0616	0.5201	0.1761	0.1460	0.0963	[-3.4707,2.3367]
Seventh fold	0.3916	0.3285	0.1119	0.1028	0.0652	[-1.6284,3.2861]
Eighth fold	0.4491	0.3457	0.0882	0.0714	0.0456	[-2.7480,1.4053]
Ninth fold	0.3871	0.3397	0.1145	0.0940	0.0646	[-3.2667,1.6763]
Tenth fold	0.3921	0.3313	0.1144	0.0951	0.0671	[-3.5077,1.6225]

Thus, for the boundary $[a_i,b_i]$ of a predicted transmembrane protein interval, there is a fuzzy interval $[c,d]$. So, a new interval $[a_i+c,a_i,b_i,b_i+d]$ can be created. Because there are five kinds of different transmembrane protein prediction algorithm, five new transmembrane intervals $[a_1+c,a_1,b_1,b_1+d]$. $[a_2+c,a_2,b_2,b_2+d]$. $[a_3+c,a_3,b_3,b_3+d]$. $[a_4+c,a_4,b_4,b_4+d]$. $[a_5+c,a_5,b_5,b_5+d]$ will be established. The final interval of the transmembrane protein $[a+c,a,b,b+d]$ can be obtained by using fusion algorithm. Obviously, the boundary of the interval of the transmembrane protein is in line with a trapezoidal distribution of the F distribution in the fuzzy theory. Its membership functions are as follows:

$$A_{(x)}=\begin{cases}0 & x\le a+c\\ (x-a-c)/(a-a-c) & a+c<x<a\\ 1 & a\le x\le b\\ (b+d-x)/(b+d-b) & b<x<b+d\\ 0 & x\ge b+d\end{cases}$$

When an appropriate threshold is determined, it can go fuzzy. In this paper, the threshold is set to 0.5, and the helix direction of the transmembrane depends on the majority of the results of the five kinds of prediction algorithms.

4.Tests and results analysis

4.1 The topology prediction accuracy of transmembrane protein sequence

If the transmembrane region of a protein sequence and their transmembrane directions are predicted correctly, the topology prediction of the entire protein sequence is correct. The topology prediction accuracy of the entire transmembrane protein sequence is expressed as follows.

$$P_T = N_{Tcor} / N_P$$

In the equation, N_P indicates the total number of membrane protein, and N_{Tcor} indicates the number of membrane protein topology is correctly predicted.

4.2 Performance evaluation of Test set

Table 4-1 shows the results of each test set, the algorithm described in this paper and is given the maximum value in indicator C and Q_P , indicating that the prediction of this algorithm is effective. Although the indicator M of the prediction algorithm ranked the second after PRODIV-TMHMM, simply compare the exact number of transmembrane domains whose prediction are accurate, you can find only three gap. Meantime, Table 4-1 shows the index value of the algorithm is the highest.

Table 4-1 Prediction accuracy of transmembrane regions and topology

Algorithm	N_{cor}	N_{obs}	N_{prd}	M	C	Q_P	N_P	N_{Tcor}	P_T
SCAMPI-seq	441	485	465	90.93%	94.84%	92.86%	124	77	62.10%
SCAMPI-msa	451	485	474	92.98%	95.15%	94.16%	124	80	64.52%
PRODIV-TMHMM	454	485	492	93.60%	92.28%	93.04%	124	76	61.29%
PRO-TMHMM	450	485	473	92.78%	95.14%	93.95%	124	74	59.68%
OCTOPUS	451	485	472	92.89%	95.55%	94.21%	124	79	63.71%
Proposed	451	485	470	92.99%	95.96%	94.46%	124	81	65.32%

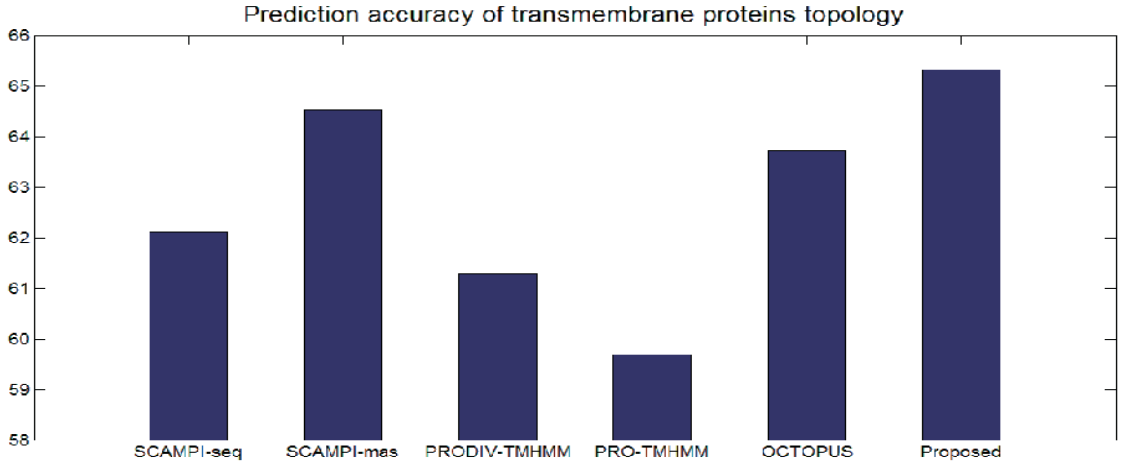

Figure 4.1 Prediction accuracy of transmembrane regions and topology

4.3 Analysis of the results

10 groups of transmembrane protein range are used for training as the training set, the weights of the five kinds of algorithms given by procedures are 0.41,0.5,0.08,0.01,0 and the fuzzy interval is [-0.58,2.16]. To predict the five

transmembrane region with the weights of the five algorithms and the fuzzy interval, the results are shown in Table 4-2.

Table 4-2 Comparing the predicted with the real

Serial number	Predicted value	Real value
1	[27.31,47.39]	[15,45]
2	[79.71,99.71]	[77,103]
3	[83.37,103.37]	[84,111]
4	[51.77,71.77]	[49,67]
5	[135.87,155.87]	[131,154]

In the evaluation, it is generally believed that as long as there are nine residues coincidence between the predicted and known transmembrane region, this forecast is correct.

It can be seen the remaining four groups have predicted accurately except group 1. The inaccuracy of group 1 is caused by the inaccuracy of the results given by the five basic prediction algorithms. Thus the correctness of the algorithm can be seen.

CONCLUSION

In this algorithm, the author takes the boundary of the transmembrane interval of the transmembrane protein as a fuzzy area, and uses the five prediction algorithms as the basis for integration and learning, then applies the results of learning to prediction. Because of the different principles of the five kinds of prediction methods and their complementarity to certain degree, higher prediction performance can be obtained by putting the predicted results of the different methods for effective integration, while the final results of this algorithm also proves this point exactly. Another characteristic of this algorithm is the introduction of fuzzy theory. This algorithm uses known trapezoidal membership functions to set different membership degree threshold to get corresponding predictions, thereby improving the flexibility of the forecast

REFERENCES

[1]Y. Deng, Q. Liu, X. Li, *Acta Chimica Sinica*, **2004**,62(19):1968~1972
[2]G. von Heijne, *Journal of Molecular Biology*, **1992**,225(2):487~494
[3]M. G. Claros, G. von Heijne, *Computer Applications in the Biosciences*, **1994**,10(6):685~686
[4]D. T. Jones, W. R. Taylor, J. M. Thornton, *Biochemistry*, **1994**,33(10):3038~3049
[5]B. Rost, R. Casadio, P. Fariselli, Refining neural network predictions for helical transmembrane proteins by dynamic programming, Proceedings of the 4th International Conference on Intelligent Systems for Molecular Biology, **1996**,:192~200
[6]X. Deng, D. Wei, Y. Li, Y. Zhang, B. Kang and Y. Deng, *Journal of Information and Computational Science*, **2011**,9(2):361~368.
[7]L. A. Zadeh, Fuzzy sets, *Information and Control*, **1965**,8(3):338~353
[8]Y. Deng, *Cybernetics And Systems*, **2011**,42(4):246~263
[9]Y. Deng, Y. Chen, Y. Zhang, and S. Mahadevan, *Applied Soft Computing*, **2012**,12(3):1231~1237
[10]Y. Deng and F. T. S. Chan, *Expert Systems With Applications*, **2011**,38(8):9854~9861
[11]Y. Deng, R. Sadiq, W. Jiang and S. Tesfamariam, *Expert Systems With Applications*, **2011**, 38 (12):15438~15446
[12]D. Wei, Y. Deng, Y. Li, Y. Zhang and S. Tang, *ICIC Express Letters*, Part B: Applications,**2012**,3(1):83~90
[13]R. R. Yager, *Fuzzy Sets and Systems*, **2003**,137(1):59~69
[14R. R. Yager, *IEEE Transactions on Systems Man and Cybernetics*, **1988**,18(1):183~190