Journal of Biomolecular Structure & Dynamics, ISSN 0739-1102
Volume 24, Issue Number 4, (2007)
©Adenine Press (2007)

Study and Prediction of Secondary Structure for Membrane Proteins

http://www.jbsdonline.com

Svetlana R. Amirova[1,*]
Juri V. Milchevsky[2]
Ivan V. Filatov[3]
Natalia G. Esipova[2]
Vladimir G. Tumanyan[2]

[1]School of Computing and Mathematics
University of Keele
Staffordshire, UK ST5 5BG.
[2]Engelhardt Institute of Molecular Biology
Russian Academy of Sciences
ul. Vavilova 32, Moscow, 119991 Russia
[3]Moscow Institute of Physics
and Technology
State University, Institutsky per. 9
Dolgoprudny, 141700, Russia

Abstract

In this paper we present a novel approach to membrane protein secondary structure prediction based on the statistical stepwise discriminant analysis method. A new aspect of our approach is the possibility to derive physical-chemical properties that may affect the formation of membrane protein secondary structure. The certain physical-chemical properties of protein chains can be used to clarify the formation of the secondary structure types under consideration. Another aspect of our approach is that the results of multiple sequence alignment, or the other kinds of sequence alignment, are not used in the frame of the method. Using our approach, we predicted the formation of three main secondary structure types (α-helix, β-structure and coil) with high accuracy, that is $Q_3 = 76\%$. Predicting the formation of α-helix and non-α-helix states we reached the accuracy which was measured as $Q_2 = 86\%$. Also we have identified certain protein chain properties that affect the formation of membrane protein secondary structure. These protein properties include hydrophobic properties of amino acid residues, presence of Gly, Ala and Val amino acids, and the location of protein chain end.

Key words: Discriminant analysis; Physical-chemical protein properties; and Formation of the protein secondary structure.

Introduction

Despite the fact that membrane proteins and their wide variety of functions are of vital importance for living cell there are few data describing their spatial structure. This void in our knowledge may even be expanded as the difficulty of protein crystallization in membrane surroundings. This complicated task very few researchers are willing to undertake. Also the majority of membrane proteins are not water-soluble, making their crystallization even more problematic. In sum, membrane protein structures make up less than 5% of all spatial protein structures in Protein Data Bank (1).

A transmembrane helix is a very common component of membrane protein secondary structure (2) and many prediction methods identify and predict only the location of the transmembrane helix. However, little is known about factors and forces that stipulate the formation of various secondary structure types in membrane proteins (3).

Many transmembrane helix prediction algorithms use the following features: I) transmembrane helices are predominantly non-polar; II) their length varies from 12 to 35 residues; III) the natural topology of a transmembrane domain is an alteration of a helix – internal loop – helix – external loop – helix structure; IV) globular regions between membrane helices are usually shorter than 60 residues; V) internal non-membrane regions are "more positively charged" than the external nonmembrane regions, that is known as the "positive external rule" (3).

[*]Phone: +44 (0)1782 583075
Fax: +44 (0)1782 584268
Email: s.amirova@epsam.keele.ac.uk

One cannot expect accurate and reliable prediction when the algorithms are based only on fast searches of homologous structures (4). Often these transmembrane prediction methods are not very effective because of the absence of suitable homologous structures or the lack of accurate homolog identification. Furthermore, there are about 30 membrane proteins for which three-dimensional homologous structures are not known (5).

There are plenty of approaches to transmembrane helix prediction from amino acid sequence (6-9). Some of them are based on multiple sequence alignment (6, 9) while the others involve transmembrane helix energy, location of charges, main features of helix folding (3), and an influence of membrane surroundings (7, 8). Undoubtedly, these methods are important in spatial structure prediction and also in the study of membrane protein folding. Molecular mechanics was applied for predicting breaks in transmembrane helices (10). Most importantly, molecular mechanics commonly is useful method for α-helix prediction (11, 12).

However, our approach to secondary structure prediction for membrane protein is not limited by helix prediction only. In addition to helix prediction, we identified the β-strand location in membrane proteins. The investigation of three dimensional structure of β-sheet proteins was carried out by various authors (13). Furthermore, using our new approach we derived certain physical-chemical protein properties affecting the formation of secondary structure in membrane proteins. Consequently, we try to explain the formation of secondary structure at concrete site using corresponding physical-chemical protein properties

Materials and Methods

Our approach to membrane protein secondary structure prediction was based on a statistical method called a stepwise discriminant analysis (14). The stepwise discriminant analysis is based on classes and predictor variables. In our approach classes were secondary structure types and predictor variables were numerous formalized protein properties. We developed a computer program that can operate with a huge number of predictor variables (up to 1000). Numerous predictor variables were included in order to take into consideration a lot of the formalized information about a protein. Also, at the last stage of the algorithm we implemented an improvement of prediction results using the profile of probability to find the certain set residues at the nearest sequence sites. This procedure can raise the prediction accuracy.

Secondary structure prediction based on statistical method stepwise discriminant analysis (14) consists of three main stages. At the first stage classes are set and the predictor variables are selected. At the second stage the predicted secondary structure is derived for every amino acid residue in the protein sequence. Finally, using the profile of probability at the nearest sequence sites an improvement of the prediction result is performed.

Training and Testing Protein Sets

For testing the method we collected a protein set which consisted of 265 membrane proteins form November 2004 release of Protein Data Bank.

For training our algorithm we used a jackknife test, which is also called a live-one-out test (14, 15). In a jackknife test the structural type of each residue was predicted by the rules derived using all other residues except the one that was being predicted. In our method we consecutively exclude groups of ten proteins that were being currently predicted. After the first group is selected, the prediction rules were derived from the other proteins. Then a different set of ten proteins was selected and the process was repeated until the predictions were made for all proteins in the dataset. This procedure dramatically reduces calculation time. This is especially

important because operation with a group of one or several residues rises the time of calculation while the accuracy of results remains practically on the same level. To summarize, we divided the protein set into 27 groups to be predicted. Every group consisted of ten proteins except the last one that consisted of five proteins. Prediction rules for the current group were derived using all the proteins except the proteins that were included in the current group.

Setting of Classes and Selection of Variables in Discriminant Analysis

Discriminant analysis method is used to obtain discriminant functions. Discriminant functions are needed to discriminate two or more classes and to select those of numerous predictor variables that are statistically significant at discriminating between classes.

In our method we set classes according to the DSSP-classification (16) of secondary structure. Class No. 1 corresponds to the DSSP designation in summary column H, G, and I, which represent helix fragments. Likewise, class No. 2 corresponds to the DSSP designation in summary column E and B, which represent beta-structure fragments. Finally, class No. 3 corresponds to the DSSP designation in summary column C, which represents coil fragments.

For the prediction model of two states – helix and non-helix states – we set two classes in a similar manner. Class No. 1 corresponds to the DSSP designation in summary column H, G, and I, and class No. 2 corresponds to the DSSP designation in summary column C.

During the selection of predictor variables, protein sequence was divided into segments. Each segment consisted of several amino acid residues and included less than ten residues. This procedure was done in order to take into consideration protein local interactions and to obtain new predictor variables changing the length of segment. As predictor variables we considered protein sequence and physical-chemical properties of current amino acid residue. These properties were mass, charge, hydrophobicity, polarity, presence of aromatic residues, presence of prolin, the localization of C- and N-chain terminus, Fourier amplitude of hydrophobic property in current segment. Also, we accepted as predictor variables the product of every physical-chemical protein property related to the sequence of current segment, which involves residue to be predicted. Along with products we consider as predictor variables squares and roots of physical-chemical protein property related with sequence of current segment. Changing the size of the segment we obtain new predictor variables. In addition, we consider as predictor variables protein properties related with whole chain sequence and various elementary mathematical functions of them. For instance the predictor variable “protein mass logarithm” was obtained in such way.

Secondary Structure Prediction for Every Amino Acid Residue

At the core of a discriminant analysis method is a stepwise procedure. To pass from one step to the next one should include or exclude a predictor variable. For every variable there is a unique succession of steps. The final aim of a stepwise procedure is to obtain discriminant functions which are used in the discrimination of classes after predictor variables $x_1 \ldots x_p$ are selected.

$$d = a + b_1 \cdot x_1 + \ldots\ldots\ldots + b_p \cdot x_p$$

The main feature of a discriminant function d_i is that its values for every considered class differ as much as possible. After appropriate discriminant functions are known we can derive the probability for every amino acid residue to be at a state corresponding to one of three considered classes. An amino acid residue is consid-

ered at a definite structure state corresponding to one of the considered classes if the probability for every amino acid residue to be at this state is maximal.

For evaluation of prediction accuracy we used the jackknife test (14). We designated a 3 × 3 matrix of prediction accuracy. The matrix element a_{ij} was the percent of residues observed at state i and predicted at state j, i,j ≡ (class No. 1, class No. 2, class No. 3). Based on value of Q_i i, ≡ (class No. 1, class No. 2, class No. 3), and Q_3 we estimated the percent of residues predicted correctly at every of three considered structural state (Q_i) and for three states (Q_3). Furthermore, based on statistical significance calculated using Fisher statistics we derive the contribution of every predictor variable with respect to the discrimination of classes and prediction of structural state for amino acid residue. Comprising statistical significance of every predictor variable we estimate its relative contribution to secondary structure prediction.

Improvement of Prediction Using Probability Profile

After the predicted structure type corresponding to one of three considered classes was obtained for every amino acid residue we improve the accuracy of the prediction. In neural networks methods this procedure is similar to a second layer of consideration. In our method, the previously predicted secondary structure was used as a predictor variable. As a result of using the profile of probability at nearest sequence sites an improvement of predicted structure is performed and therefore the accuracy of secondary structure prediction rises. This improvement of predicted structure can be illustrated by the following examples. Let us consider impossible the prediction of beta-structure of one residue among long helix segment or prediction of helix state for a proline residue. Using the profile of probability at nearest sequence sites this type of false prediction can be eliminated.

Results and Discussion

Results of secondary structure prediction for membrane proteins are shown in Table I. Table I represents a 3 × 3 matrix of prediction accuracy and every matrix element a_{ij} is the percent of residues observed at state i and predicted at state j, i,j ≡ ("H", "E", "C"). In the 3 × 3 matrix the diagonal elements, which are shown in bold, represent the value Q_{3i}, which is the percent of residues predicted correctly at each of the three considered structural states. The fraction of correctly predicted helix residues was Q_{3H} = 78%, for beta-strands the corresponding value was Q_{3E} = 69% and, finally the value Q_{3C} = 78% was observed for coil residues. Calculated from this matrix of accuracy, the percent of residues predicted correctly at three considered structural state is value Q_3 = 76%.

Table I
Matrix of accuracy for membrane protein secondary structure prediction (%).

Observed secondary structure	Predicted secondary structure		
	coil class "C"	α-helix class "H"	β-structure class "E"
Coil class "C"	**78**	11	11
α-helix class "H"	12	**78**	10
β-structure class "E"	18	13	**69**

Results of secondary structure prediction using model of two stages α-helix and non-α-helix in membrane proteins are shown in Table II. Table II represents a 2 × 2 matrix of prediction accuracy and every matrix element a_{ij} is the percent of residues observed at state i and predicted at state j, i,j ≡ ("H", "non-H"). In the 2 × 2 matrix the diagonal elements, which are shown in bold, represent the value Q_{2i}, which is the percent of residues predicted correctly at each of the two considered structural states. The fraction of correctly predicted α-helix states was Q_H = 82%, for non-α-helix states the corresponding value was Q_{nonH} = 88%. The value 12% represents the percent of residues for which the observed α-helix was pre-

Table II
Prediction of α-helixes in membrane proteins (%)

Observed secondary structure	Predicted secondary structure	
	Non-α - helix class No. 2	α-helix class No. 1
Non-α-helix	**88**	12
α-helix "H"	18	**82**

dicted as a non-α-helix and the value 18% represents the percent of residues for which the observed non-α-helix was predicted as an α-helix. Calculated from the matrix of accuracy, the percent of residues predicted correctly at two considered structural state was the value $Q_2 = 86\%$.

We have tried to compare our prediction accuracy with other methods that are used to predict membrane protein secondary structure. First of all, we have noted that the majority of the methods are applied for prediction transmembrane helix only and few methods offer the prediction of beta-structure and coil regions in membrane proteins. Second, most authors used completely different training and testing protein data set to calculate the accuracy of their method. Therefore, it is difficult to perform direct comparison of prediction methods because of various used protein sets and different accuracy criteria used by many authors.

One of the most widely used methods to predict secondary structure in membrane proteins is PHDhtm method by Rost, B. and Sander, C. (17). A combination of three levels of networks that authors used in their method results in an overall three-state accuracy of 70.8% for globular proteins and if membrane protein chains are included in the evaluation, the overall accuracy reduces to 70.2%. The prediction is well balanced between alpha-helix, beta-strand, and loop: 65% of the observed strand residues are predicted correctly. For half of the residues predicted with a high level of reliability the overall accuracy increases to better than 82%. So our approach demonstrates better level of accuracy providing $Q_2 = 86\%$ for α-helix prediction.

The popular method to predict secondary structure in membrane proteins is TOPPRED (18). This method based on the "positive inside" rule (19) and uses extra information in the form of the distribution of positively charged residues. This method was very successful for predicting the topology of bacterial inner membrane proteins. A straightforward implementation with no attempts at optimization predicts the correct topology for 23 out of 24 inner membrane proteins with experimentally determined topologies, and correctly identifies 135 transmembrane segments with only one over prediction. Our approach offers the prediction of secondary structure for each residue of the current membrane protein. Also using our method one can consider prediction of both three main secondary structure types and α-helix only.

Another popular method to predict secondary structure in membrane proteins is HMMTOP by Tusnady, G. E. and Simon, I. (20). The authors tested the prediction accuracy on 158 proteins. The method successfully predicted all the transmembrane segments in 143 proteins out of the 158, and for 135 of these proteins both the membrane spanning regions and the topologies were predicted correctly. But we noted that our testing set was larger and consisted of 265 membrane proteins and we calculated accuracy using each residue from testing protein set.

The SOSUI method by T. Hirokawa, *et al*. (21), which gives not only secondary structure prediction for membrane proteins but also a classification of secondary structure in membrane proteins. The method offers prediction for transmembrane helix only and the accuracy of this prediction was 97% but the prediction of other secondary structure types such as β-structure and coil was not taken into consideration in the method.

The rather accurate method to predict transmembrane segments is PRED-TMR by C. Pasquier, *et al*. (22). This method used three main criteria to evaluate the prediction accuracy for transmembrane helixes, one of them is Q value similar to our Q_2 value in prediction of α-helix. The results of the test on this set of 101 non-homologous transmembrane proteins gave an average Q value of 88.83%. The authors achieved slightly lower accuracy Q value of 86.14% using SwissProt, release 35, contains 9392 transmembrane sequences with a total of 40,672 transmembrane regions. Again β-structure and coil was not taken into consideration in the method.

A number of algorithms designed to identify transmembrane helices in the primary amino acid structure have been developed, and current methods can identify around 90-95% of all true transmembrane segments with an over-prediction rate of only a few percent. The best results have so far been obtained when multiply aligned sequences can be analyzed; however, in many cases there are no homologs in the database and improvements in single-sequence prediction performance are thus important.

In our approach we have some novel aspects comparable to other methods. First, we offered the prediction not only of α-helix but of β-structure and coil regions in membrane proteins. Second, we did not use any kind of sequence alignment in the frame of the method. And finally, we identified protein chain properties that affected the secondary structure of membrane proteins.

Protein chain properties dramatically effected the formation of considered structural types are shown in Table III. Every protein property goes with corresponding statistical significance calculated using Fisher statistics in the discriminant analysis.

Table III

Protein chain properties that effected the secondary structure prediction in membrane proteins (only the affected properties are shown).

Protein properties considerably affected by secondary structure prediction	Statistical significance of corresponding predictor variable (using Fisher statistic)
Presence of tiny residues Ala, Gly, Val on segment of seven residues in length the center of which is the firth residue on the right from the current residue	2406.461
End of protein chain on the segment of three residues on the left from the current residue	1695.379
Hydrophobicity of three residues segment the center of which is the second residue on the right from the current residue	1637.635
Presence of amino acid proline on the segment of seven residues the center of which is the third residue on the left from the current residue	686.966
Logarithm of entire protein chain mass	192.807

Among the most important protein properties appeared to be the presence of tiny residues Ala, Gly, and Val, the location of protein chain end and hydrophobic property of certain sets of amino acid residues. The point is that the main feature of secondary structure formation in membrane proteins is a considerable influence of local segment hydrophobic property and local interaction including tiny residues and location of chain end. In addition, from Table III one can see that the global protein property "Logarithm of entire protein chain mass" also contributes to secondary structure formation in membrane proteins.

Conclusion

Using the new approach based on the method of a discriminant analysis we predicted the formation of three main secondary structure types (α-helix, β-structure, and coil) in membrane proteins with a rather high accuracy (Q_3 = 76%). Furthermore, for each structural type the accuracy was Q_{3H} = 78% for the helix state, Q_{3E} = 69%

for the beta-strand state, and Q_{3C} = 78% finally for the coil state. For prediction of the formation of α-helix and non-α-helix states we reached the accuracy of Q_2 = 86%. In addition the value of α-helix correct prediction was Q_H = 82%.

At the same time, we have identified certain protein properties that effect the formation of membrane protein secondary structure and derived the relative contribution of each property. These protein properties include hydrophobic properties of amino acid residues, presence of Gly, Ala, and Val amino acids and the location of the protein chain end. We have shown that not only local but global protein properties effect the secondary structure formation in membrane proteins. Thus, we try to explain some features of formation in membrane proteins.

The prediction method developed is available on http://prophet.imb.ac.ru/

Acknowledgments

This work was supported by Russian Fund of Basic Researches (projects No 05-04-4962 and No 03-04-4917), by Presidium RAS (the program "Molecular and Cell Biology"), and by Ministry of Education and Science (contract No 02.434.11.1008).

We would like to thank F. Kondrashov for critical reading of the manuscript and valuable comments.

References and Footnotes

1. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., Bourne, P. E. *Nucleic Acids Research 28*, 235-242 (2000).
2. Von Heijne, G. *Prog Biophys Mol Biol 66*, 113-139 (1996).
3. Von Heijne, G. *EMBO J 5*, 3021-3027 (1986).
4. Thompson, J. D., Higgins, D. G., Gibson, T. J., Clustal, W. *Nucleic Acids Res 22*, 4673-4680 (1994).
5. Saier M. H. Jr. *J Cell Biochem Suppl 32-33*, 84-94 (1999).
6. Von Heijne, G., Gavel, Y. *Eur J Biochem 174*, 671-678 (1988).
7. Jayasinghe, D., Hristova, K., White, S. H. *J Mol Biol 312*, 927-934 (2001).
8. Jones, D. T., Taylor, W. R., Tornton, J. M. *Biochemistry 33*, 3038-3049 (1994).
9. Moller, S., Croning, M. D., Apweiler, R. *Bioinformatics 17*, 646-653 (2001).
10. Nikiforovich, G. V. *Prot Engineering 11*, 279-283 (1998).
11. Kilosanidze, G. T., Kutsenko, A. S., Esipova, N. G., Tumanyan, V. G. *FEBS Lett 510*, 13-16 (2002).
12. Kilosanidze, G. T., Kutsenko, A. S., Esipova, N. G., Tumanyan, V. G. *Prot Sci 13*, 351-357 (2004).
13. Nikiforovich, G. V., Frieden, C. L. *Proc Natl Acad Sci 99*, 10388-10393 (2002).
14. Mardia, K. V., Kent, J. T., Biddy, J. M. *Multivariate Analysis*. Academic Press, London (1979).
15. Klein, P. *Biochem. Biophys. Acta. 874*, 205-215 (1986).
16. Kabsch, W., Sander, C. *Biopolymers 22*, 2577-2637 (1983).
17. Rost, B., Sander, C. *J Mol Biol 232*, 584-599 (1993).
18. Claros, M. G., von Heijne, G. *CABIOS 10*, 685-686 (1994).
19. Von Heijne, G. *J Mol Biol 225*, 487-494 (1992).
20. Tusnády, G. E., Simon, I. *J Mol Biol 283*, 489-506 (1998).
21. Hirokawa, B.-C., Mitaku. *Bioinformatics 14*, 378-379 (1998).
22. Pasquier, C., Promponas, V. J., Palaios, G. A., Hamodrakas, J. S., Hamodrakas, S. J. *Protein Eng 12*, 381-385 (1999).

Date Received: August 23, 2005

Communicated by the Editor Valery I Ivanov