ISSN 0026-8933, Molecular Biology, 2006, Vol. 40, No. 3, pp. 481–485. © Pleiades Publishing, Inc., 2006.
Original Russian Text © R.A. Sutormin, A.A. Mironov, 2006, published in Molekulyarnaya Biologiya, 2006, Vol. 40, No. 3, pp. 541–545.

MATHEMATICAL AND SYSTEM BIOLOGY

UDC 577.1

Membrane Profile-Based Probabilistic Method for Predicting Transmembrane Segments via Multiple Protein Sequence Alignment

R. A. Sutormin[a] and A. A. Mironov[a, b, c]

[a] *State Research Center GosNIIgenetika, Moscow, 117545 Russia; e-mail: sutor_ra@mail.ru*

[b] *Institute of Information Transmission Problems, Russian Academy of Sciences, Moscow, 127994 Russia*

[c] *Department of Bioengineering and Bioinformatics, Moscow State University, Moscow, 119992 Russia*

Received January 25, 2006

Abstract—Prediction of transmembrane (TM) segments of amino acid sequences of membrane proteins is a well-known and very important problem. The accuracy of its solution can be improved for approaches that do not use a homology search in an additional data bank. There is a lack of tested data in this area of research, because information on the structure of membrane proteins is scarce. In this work we created a test sample of structural alignments for membrane proteins. The TM segments of these proteins were mapped according to aligned 3D structures resolved for these proteins. A method for predicting TM segments in an alignment was developed on the basis of the forward–backward algorithm from the HMM theory. This method allows a user not only to predict TM segments, but also to create a probabilistic membrane profile, which can be employed in multiple alignment procedures taking the secondary structure of proteins into account. The method was implemented in a computer program available at http://bioinf.fbb.msu.ru/fwdbck/. It provides better results than the MEMSAT method, which is nearly the only tool predicting TM segments in multiple alignments, without a homology search.

DOI: 10.1134/S0026893306030150

Key words: membrane protein, secondary structure prediction, hidden Markov models, forward–backward algorithm, probabilistic membrane profile, test sample

INTRODUCTION

Many problems of bioinformatics include the stage of aligning the amino acid sequences of proteins [1], so the quality of alignment is often critical for sequence analysis [2]. In the case of membrane proteins, it is quite difficult to determine their 3D structures [3] and crystallographic data for such molecules are scarce. Hence, it is difficult to construct structural alignments necessary for the verification of multiple sequence alignments obtained by automatic methods [4]. On the other hand, there is good reason to believe that common alignment procedures do not ensure good results, because of the unusual and irregular amino acid composition of membrane proteins [5]. There is the opinion that information about the secondary structure of proteins should improve alignment quality [6]. In the case of membrane proteins, a consistent decision is to take the protein regions located within the membrane as such secondary-structure elements [7], since other regions, located beyond the membrane, resemble globular proteins in properties, and globular proteins can be aligned quite well [8]. Thus, it is necessary to learn how to correctly predict the transmembrane (TM) segments in a protein sequence before developing an alignment algorithm based on such predictions. Unfortunately, the quality of the methods predicting TM segments, by the primary amino acid sequence, is far from ideal [9]. There are some methods, such as PHDpsihtm (part of the PredictProtein server) [10] or MEMSAT (online mode) [11], that predict the protein structure based on the results of a search of homologous sequences in a protein data bank. In this work we focused on this problem, seeking its solution without resorting to an additional homology search. We have failed to find an approach that would predict TM segments on the basis of multiple sequence alignments, and would not require additional homology data, except for the MEMSAT method (offline mode). This method is based on choosing an appropriate TM model with the help of dynamic programming according to the preferential location of amino acid residues on the surface, within, or beyond the membrane. The method yields a

sequence mask, where each symbol informs about the location (within or beyond the membrane) of the amino acid residues from the corresponding alignment column. Since there is no certainty that the position of a TM segment is rigidly fixed (a protein molecule "breathes"; i.e., there are weak fluctuations of the chain links), it is more adequate to predict the probability of the intramembrane location of the given amino acid (probabilistic membrane profile). This probability should be rather high in the case of amino acids located within a membrane, and should smoothly decrease to zero with amino acids located on the borders of a membrane. It is possible to obtain such probabilities by constructing a hidden Markov model (HMM) for intramembrane and external regions of a protein. Using the forward–backward algorithm [12], we can calculate the probability for a given amino acid to belong to the membrane part of the model.

The objective of this work was to create a set of membrane protein clusters with the resolved 3D structures, for which it is possible to construct an adequate structural alignment, and to develop a method for constructing the membrane profile for the columns of multiple alignment.

METHODS

The procedure of constructing a membrane profile where every column of a multiple alignment has a certain probability of the intramembrane location of its amino acid residues consists in the following. The amino acid frequency profile (frequency matrix) is derived on the basis of multiple sequence alignment. For this purpose it is necessary to derive a matrix of pairwise evolutionary distances between the sequences, based on pairwise identity and using the Poisson correction procedure [13]: $d = -\log((20\max\{1.1/20, id\} - 1)/19)$, where d is the evolutionary distance and id is the fraction of alignment columns with the same amino acids. Then, a phylogenetic tree is designed using the nearest neighbor procedure [14], and a weight is assigned to every sequence via a simple and efficient method proposed by Gerstein et al. [15]. The weights have the following property. Weight $1/k$ is assigned to each of the k identical sequences, while weight 1 is assigned to each unique sequence. A frequency profile is created by averaging all sequence profiles according to their weights. To obtain a result, we used the forward–backward algorithm based on the HMM, similar to the model used by the TMHMM server [16]. This model includes the states of amino acids located in the cytoplasm and those exposed out of the cell and the two sequences of states corresponding to a protein chain crossing the membrane outward and in the reverse direction. Two groups of model states are distinguished that correspond to the membrane borders. The parameters of the model were trained with a sample of single amino acid sequences with determined TM segments, which are available from the TMHMM server web site. This server predicts TM segments in a single amino acid sequence, but cannot work with frequency matrix or multiple sequence alignment. The work of the server is based on the Witherby algorithm (see [12]), which is used to find an optimal path, but in contrast to the forward–backward algorithm cannot be used to build a probabilistic profile.

Test Sample Construction

A sample of standard multiple sequence alignments was created to test our method. We took all sequences of membrane proteins with the resolved 3D structure (442 proteins) available from the PDBTM server web site [17]. Then we constructed all pairwise alignments, using the CLUSTALW program [1]. When we found a pair of proteins with 95% or higher identity, we excluded one of them from consideration. After this we carried out clustering according to pairwise identity, using the nearest neighbor procedure [14] with the lower threshold set at 20%. When the size of a cluster exceeded 20 proteins, a cluster was divided into smaller ones by raising the lower threshold of clustering. Only clusters containing at least three proteins were taken into further consideration. We carried out multiple structural alignment of the 3D structures of proteins for every cluster, using the MAMMOTH server [18]. In the case of low alignment quality, i.e., a small number of reliable (in view of this approach) alignment columns, we rejected the most remote member of the cluster and did the alignment again.

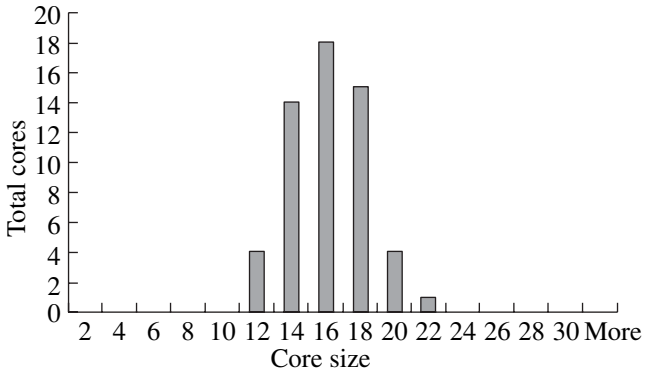
The final sample contained 11 clusters of 55 proteins. The fraction of structurally reliable alignment columns varied from 24 to 86%, averaging 63%. The size of clusters varied from three to eight proteins; the average number of proteins per cluster was five. We examined proteins from these clusters for their belonging to structural families in accordance with the SCOP [19] and CATH [20] classifications. We found a two-domain structure in one cluster, and there were some proteins containing only one of these domains. Three clusters contained proteins whose structural families have not been identified in both classification systems. Two clusters contained proteins from different families: one combined the bacteriorhodopsin (SCOP f.13.1.1) and succinate dehydrogenase/fumarate reductase (SCOP f.21.2.2) families, and the other proteins belonging to the reaction center of photosystem I (SCOP f.31.1.1) and the ATPase domain of an ABC transporter (SCOP c.37.1.12).

Construction of Reliable Membrane Marking

For every protein of every cluster we determined the TM segment using the TMDET algorithm [21],

which determines the most probable position of the membrane across a 3D structure. We allowed the 5-Å "twilight" areas at the margins of the membrane to avoid the misclassification of a protein region as a TMDET segment in the cases of the inaccurate prediction of the membrane location by this algorithm. If a sequence region was located completely within a twilight area, it was not marked as TM. We applied this marking to structural alignments to create a common TM marking (TM cores). These TM cores included only those columns of a structural alignment where all amino acid positions (except for deletions) had been marked as TM residues. According to the output information of the MAMMOTH server about the reliability of structural alignment in particular columns, all cores were divided into two classes, reliable and doubtful. The first one contained cores in which the alignment was reliable for two-thirds of the columns according to the MAMMOTH server, and whose length was not less than five columns. Doubtful cores were rejected.

As a result we obtained 56 TM cores. The number of cores per alignment varied from 1–12 and five on average. In addition in three of the alignments, we rejected five doubtful TM cores where the portion of reliable (in relation to the structural alignment) columns was less than 60%. The size distribution of the TM cores is shown in the figure.



Size distribution of TM cores.

Methods for Predicting the Membrane Marking from the Alignment

We examined the following methods predicting the TM segments: MEMSAT; FWDBCK, based on the above procedure of constructing the probabilistic TM profile; and a procedure of averaging the results obtained using the HMMTOP server [22] over a protein alignment (hereafter referred to as HMMTOP averaged). We used amino acid frequency profiles of the alignments as input parameters for the MEMSAT server, taking the sequence weights into account, disregarding deletions. FWDBCK marking of TM segments was carried out as follows. Columns with a probability of the intramembrane location equal to 0.8 or more were considered TM columns. However, if there were less than five adjacent TM columns, they were not referred to this category.

The HMMTOP averaging procedure was organized the following way. For every alignment, the TM segments predicted by the HMMTOP server were mapped on every aligned amino acid sequence. Columns were considered as TM columns if at least two-thirds of the deletion-free positions had been marked as TM residues. If there were less than five adjacent TM columns, they were not referred to this category.

Quality Assessment for the HMMTOP Procedure

We examined the quality of the HMMTOP procedure for every protein sequence of every cluster to make sure that the methods predicting TM segments, based on alignment, are better than those operating with only a single sequence. For each amino acid sequence, we narrowed the information about the reliability of columns in the structural alignment of the corresponding cluster. To do this we rejected the columns containing deletions from the given sequence. In the same manner, we positioned the TM cores on the amino acid sequence; so we narrowed the marking made for the whole alignment. Then we applied a filter to this marking and the one predicted by the HMMTOP procedure. The filter was similar to the one above and allowed us to ignore the TM segments and cores with a small length or a low degree of overlapping with the reliability mask. The result is shown in the table in the HMMTOP(orig.) row.

Assessment of the Prediction Quality

Before assessing the quality of the TM marking predicted by any method, it is necessary to reject all

Quality of the predictions of TM segments with different methods

Method	Quality_five[1]	Quality_half[2]
MEMSAT	0.964	0.964
FWDBCK	0.977	0.966
HMMTOP (aver.)	0.934	0.934
HMMTOP (orig.)	0.916	0.914

Notes: [1] The prediction quality represents a ratio of the number of cores having at least five columns covered by any predicted TM segment to the maximum of the total number of cores and the total number of predicted segments.

[2] The quality of a prediction represents a ratio of the number of cores having at least 50% of columns covered by any predicted TM segment to the maximum of the total number of cores and the total number of predicted segments.

doubtful TM segments. We considered a region as reliable, if two-thirds of its columns had a reliable structural alignment according to the MAMMOTH server, and the length of the segment was at least five columns. Two quality parameters were calculated for each prediction method and cluster. The first parameter (quality_five) represents a ratio of the number of cores having at least five columns covered by any predicted TM segment to the maximum of the total number of cores and the total number of predicted segments. The second parameter (quality_half) represents a ratio of the number of cores having at least 50% of the columns covered by a predicted TM segment to the maximum of the total number of cores and the total number of predicted segments. As evident from the table, the best results were obtained using the FWDBCK method.

RESULTS

Today, bioinformatics lacks data on membrane proteins suitable for verifying the quality of different methods for an automatic prediction of TM segments and construction of multiple amino acid sequence alignments. Most of the alignment parts presented in the membrane protein section of Balibase [4] lack TM-segment marking, which might have been obtained through analyzing the resolved 3D structures. In addition this section lacks the highlighting of columns whose alignment is reliable according to the structural alignment method.

In this work we created a sample of membrane protein clusters. For every cluster a structural multiple alignment was constructed and TM cores (i.e., groups of columns belonging to the intramembrane class according to the structure of each protein of the cluster) were positioned. Although the average length of cores was 15.5 residues, less than the average length commonly accepted for a TM segment (21 residues), the cores did not contain any doubtful columns. In addition, we determined the columns that are reliable according to the structural alignment procedure. Thus, in spite of the small size, this sample can be used to estimate the quality of methods predicting the TM segments or constructing multiple sequence alignments.

On the other hand, we developed a method for constructing a probabilistic membrane profile. The adequacy of the method was tested through analyzing the TM-segment prediction obtained with its use (see FWDBCK in the table). The quality of this prediction proved to be slightly better than with the most precise methods avoiding a homology search in an additional data bank.

Such profiles can also be used for constructing multiple amino acid sequence alignments of membrane proteins. If the alignment procedure is progressive, then, at every step, a merging of two subalignment profiles into one makes it possible to improve the resulting alignment by varying some column parameters, such as the substitution matrix or penalties for the opening and extension of gaps. Such variations should depend on the probability of the intramembrane location of amino acids of the given column.

In addition, we have developed an Internet server that allows a user to obtain a probabilistic membrane profile for any alignment.

The server and the test sample are available at http://bioinf.fbb.msu.ru/fwdbck/.

ACKNOWLEDGMENTS

This work was supported by the programs Molecular and Cell Biology and The Origin and Evolution of the Biosphere of the Russian Academy of Sciences, the Howard Hughes Medical Institute (grant no. 55000309), and the Russian Foundation for Basic Research (project no. 05-04-48759).

REFERENCES

1. Thompson J.D., Higgins D.G., Gibson T.J. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*. **22**, 4673–4680.
2. Jaroszewski L., Li W., Godzik A. 2002. In search for more accurate alignments in the twilight zone. *Protein Sci*. **11**, 1702–1713.
3. Zhang H., Cramer W.A. 2005. Problems in obtaining diffraction-quality crystals of hetero-oligomeric integral membrane proteins. *J. Struct. Funct. Genomics*. **6**, 219–223.
4. Bahr A., Thompson J.D., Thierry J.C., Poch O. 2001. BAliBASE (Benchmark Alignment dataBASE): Enhancements for repeats, transmembrane sequences, and circular permutations. *Nucleic Acids Res*. **29**, 323–326.
5. Sutormin R.A., Rakhmaninova A.B., Gelfand M.S. 2003. BATMAS30: Amino acid substitution matrix for alignment of bacterial transporters. *Proteins*. **51**, 85–95.
6. Heringa J. 1999. Two strategies for sequence comparison: Profile-preprocessed and secondary structure-induced multiple alignment. *Comput. Chem*. **23**, 341–364.
7. Ng P.C., Henikoff J.G., Henikoff S. 2000. PHAT: A transmembrane-specific substitution matrix. Predicted hydrophobic and transmembrane. *Bioinformatics*. **16**, 760–766.
8. Do C.B., Mahabhashyam M.S., Brudno M., Batzoglou S. 2005. ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Res*. **15**, 330–340.
9. Chen C.P., Kernytsky A., Rost B. 2002. Transmembrane helix predictions revisited. *Protein Sci*. **11**, 2774–2791.
10. Rost B., Liu J. 2003. The PredictProtein server. *Nucleic Acids Res*. **31**, 3300–3304.

11. Jones D.T. 1998. Do transmembrane protein superfolds exist? *FEBS Lett*. **423**, 281–285.

12. Krogh A., Mian I.S., Haussler D. 1994. A hidden Markov model that finds genes in *E. coli* DNA. *Nucleic Acids Res*. **22**, 4768–4778.

13. Zuckerkandl E., Pauling L. 1965. Molecules as documents of evolutionary history. *J. Theor. Biol*. **8**, 357–366.

14. Saitou N., Nei M. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol*. **4**, 406–425.

15. Gerstein M., Sonnhammer E.L., Chothia C. 1994. Volume changes in protein evolution. *J. Mol. Biol*. **236**, 1067–1078.

16. Sonnhammer E.L., von Heijne G., Krogh A. 1998. A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol*. **6**, 175–182.

17. Tusnady G.E., Dosztanyi Z., Simon I. 2005. PDB_TM: Selection and membrane localization of transmembrane proteins in the protein data bank. *Nucleic Acids Res*. **33**, 275–278.

18. Lupyan D., Leo-Macias A., Ortiz A.R. 2005. A new progressive-iterative algorithm for multiple structure alignment. *Bioinformatics*. **21**, 3255–3263.

19. Murzin A.G., Brenner S.E., Hubbard T., Chothia C. 1995. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol*. **247**, 536–540.

20. Orengo C.A., Michie A.D., Jones S., Jones D.T., Swindells M.B., Thornton J.M. 1997. CATH: A hierarchic classification of protein domain structures. *Structure*. **5**, 1093–1108.

21. Tusnady G.E., Dosztanyi Z., Simon I. 2005. TMDET: Web server for detecting transmembrane regions of proteins by using their 3D coordinates. *Bioinformatics*. **21**, 1276–1277.

22. Tusnady G.E., Simon I. 1998. Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *J. Mol. Biol*. **283**, 489–506.