A novel tool for the prediction of transmembrane protein topology based on a statistical analysis of the SwissProt database: the OrienTM algorithm

T D Liakopoulos, Claude Pasquier, S J Hamodrakas

Protein Engineering vol.14 no.6 pp.387–390, 2001

COMMUNICATION

A novel tool for the prediction of transmembrane protein topology based on a statistical analysis of the SwissProt database: the OrienTM algorithm

Theodore D.Liakopoulos, Claude Pasquier and Stavros J.Hamodrakas[1]

Department of Cell Biology and Biophysics, Faculty of Biology, University of Athens, Panepistimiopolis, Athens 157 01, Greece

[1]To whom correspondence should be addressed. E-mail: shamodr@cc.uoa.gr

OrienTM is a computer software that utilizes an initial definition of transmembrane segments to predict the topology of transmembrane proteins from their sequence. It uses position-specific statistical information for amino acid residues which belong to putative non-transmembrane segments derived from statistical analysis of non-transmembrane regions of membrane proteins stored in the SwissProt database. Its accuracy compares well with that of other popular existing methods. A web-based version of OrienTM is publicly available at the address http://biophysics.biol.uoa.gr/OrienTM.

Keywords: prediction/topology/transmembrane proteins/World Wide Web

Introduction

Transmembrane proteins play important roles in cellular functions. Unfortunately, even today, it is very difficult to solve their 3-D structure by X-ray crystallography or NMR (Persson and Argos, 1994; von Heijne, 1996; Aloy *et al.*, 1997; Berman *et al.*, 2000). Thus, several successful prediction algorithms have been developed for transmembrane proteins, which not only predict transmembrane segments, but also topology, secondary structure and, sometimes, even secondary structure packing (references cited in Tusnády. and Simon, 1998; Pasquier *et al.*, 1999; Promponas *et al.*, 1999).

The successful location of transmembrane segments, of their secondary structure and the packing modes of secondary structure elements is important because they define the architecture of a transmembrane protein (von Heijne, 1996). However, equally important is the determination of topology, which defines the 'polarity' of integral membrane proteins. Topology can be determined experimentally (Jennings, 1989) or predicted by computational methods (von Heijne, 1992; Sipos and von Heijne, 1993; Jones *et al.*, 1994; Fariselli and Casadio, 1996; Rost *et al.*; 1996, Diederichs *et al.*, 1998). Some computational methods depend primarily on a series of rules derived from observation and statistical studies. The most successful among them is the 'positive inside rule' (von Heijne, 1992), which simply states that the propensity of positively charged residues (basically lysine and arginine) is higher in the non-transmembrane segments on the inner part of the cell. Other similar rules state that the positive inside bias is most visible at the N-terminal end of the sequence, also that a bias of the negative residues also exists and that a high propensity of tyrosine and tryptophan indicates the outer part of the cell (Sipos and von Heijne, 1993). Recently, a hidden Markov model was developed for topology prediction of helical transmembrane proteins (Tusnády and Simon, 1998). It is based on the hypothesis that the localization of the transmembrane segments and the topology are determined by the difference in the amino acid distributions in various structural parts of these proteins rather than by specific amino acid compositions of these parts. The authors achieved 85% prediction accuracy on a set of 158 proteins (both the topology and the transmembrane segments were predicted correctly), which they claim is higher than that found using prediction methods already available.

In this paper, we present a simple algorithm, which predicts the topology of transmembrane proteins from sequence alone, given the transmembrane segments of the protein. It utilizes position-specific parameters for residues which belong to non-transmembrane segments. These were derived from a statistical analysis of every non-transmembrane segment in a database of non-transmembrane protein regions, DB-NTMR (freely available at http://biophysics.biol.uoa.gr/DB-NTMR/), automatically derived from the Swiss-Prot database (release 35) of protein sequences (Bairoch and Apweiler, 1998). The tool was extensively tested on several test sets of sequences available and was also applied to the whole SwissProt database (release 35).

Methods

A database of non-transmembrane protein regions, DB-NTMR (freely available at http://biophysics.biol.uoa.gr/DB-NTMR/), was automatically derived from the Swiss-Prot database (release 35) of protein sequences (Bairoch and Apweiler, 1998).

Calculation of statistical parameters

A position-specific residue statistical analysis was made of every non-transmembrane segment (NTS) in DB-NTMR. Initially, four-dimensional arrays $A[2,20,4,20]$ were calculated counting the number of residues relevant to:

(i) whether the NTS exists on the inner or outer part of the cell (values: 0–1);
(ii) residue type (values: 1–20);
(iii) whether a residue is nearest to (a) the C-terminus, (b) the N-terminus of a transmembrane segment (TS) and (c) the C-terminus or (d) the N-terminus of a transmembrane protein end (values: 1–4);
(iv) residue position relative to the closest TS or molecule end mentioned above (values: 1–20).

A residue was not taken into account if its distance from the closest NTS end is larger than 20 nominal residue positions. Apparently, such residues exist in NTSs longer than 40 amino acids. Non-transmembrane segments shorter than 40 residues are divided into two equal parts if they contain an even number of residues. In such segments, the middle residue was not taken into account if the number of residues is odd.

Finally, a set of three-dimensional arrays $B[20,4,20]$ of indicators was calculated, containing the ratio of residues

which are on the inner part to those on the outer part, multiplied by the ratio of the total residue number on the outer part to the total number on the inner part:

$$B[x,y,z] = \frac{A\,[0,x,y,z]}{\sum_{i=1}^{20} A[0,i,y,z]} \Bigg/ \frac{A\,[0,x,y,z]}{\sum_{i=1}^{20} A[0,i,y,z]}$$

These arrays were transformed so that their minimum value is –1 (perfect tendency of a residue type in a particular position to be in the inner part) and their maximum value is +1 (perfect tendency to be in the outer part). The transformation is given by

$$C[x,y,z] = \frac{B[x,y,z]-1}{B[x,y,z]+1}$$

Prediction phase

Two scenarios are possible for each protein molecule to predict:

(a) the first NTS being on the inner part of the cell, the next on the outer and so on ('in' topology), and
(b) the first NTS being on the outer part and so on ('out' topology).

The NTSs of the molecule are scanned and a score is calculated by adding ('odd' NTSs) and subtracting ('even' NTSs) the respective $C[x,y,z]$ indicators for each residue. The score indicates whether this is likely to be a true or false topology scenario (positive or negative values) and how likely that is (absolute score value). The score for the other scenario is obviously the opposite number.

Complementary predictions

Adding up the indicators for each NTS separately, probabilities for the topology of that single NTS are revealed, either extracellular or cytoplasmic (positive or negative values). The same process may be carried out for each TS (calculating scores for the non-transmembrane residues flanking the TS). In this case, positive values correspond to transmebrane segments (with an N- towards C-terminal orientation) which cross the membrane from the extracellular to the cytoplasmic side, whereas negative values correspond to transmembrane segments which have the opposite orientation.

Results

The method was tested initially on a set of 72 proteins (extracted from the set of the 101 non-homologous transmembrane proteins reported by Pasquier *et al.*, 1999), with the criterion that the topology is known) and on all transmembrane proteins with known topology of the SwissProt database (release 35) (Bairoch and Apweiler, 1998). The results are shown in Table I and clearly indicate that the algorithm predicts the topology of eukaryotic transmembrane proteins with high accuracy, whereas the accuracy level for prokaryotic transmembrane proteins is lower. This observation was also noted previously and no satisfactory explanation could be found (Rost *et al.*, 1996). Most of the false predictions correspond to low absolute scores. On the other hand, most of the high absolute scores correspond to correct predictions.

It might be argued that the high scores of correct prediction were obtained because the proteins included in the test sets were also contained in the training set (which consists of all transmembrane proteins with known topology contained in SwissProt, release 35, as shown in Table I). However, this is not true. To evaluate the performance of the algorithm on new sequences not contained in the training set, we tested it additionally on all SwissProt entries new in releases 36–39 (Table I). It was found that the percentages of eukaryotic and prokaryotic proteins with topology predicted correctly (94 and 87%, respectively), are almost identical with the percentages predicted taking as test set the proteins of the training set (95 and 87%, respectively; see also Table I).

Table I. Results obtained utilizing OrienTM on several test sets: (a) on a subset of 72 transmembrane proteins with known topology of the set of 101 proteins used by Pasquier *et al.* (1999) (Set of '72'); (b) on all transmembrane proteins of known topology of the SwissProt database (release 35) (Bairoch and Apweiler, 1998); this set was also used as the 'training' set; (c) on all transmembrane proteins of known topology contained in SwissProt releases 36–39; (d) on the test set used by HMMTOP (Tusnády and Simon, 1998) and (e) on the sets presented by Möller *et al.* (2000)

Test set		No. of proteins	Topology correctly predicted by OrienTM
Set of '72'	Eukaryotic	38	37 (97%)[a]
	Prokaryotic	34	33 (97%)[a]
Subset of SwissProt (release 35) with known topology (used as 'training set')	Eukaryotic	3240	3080 (95%)
	Prokaryotic	451	392 (87%)
Subset of SwissProt (new entries in releases 36–39) with known topology	Eukaryotic	588	550 (94%)
	Prokaryotic	99	86 (87%)
HMMTOP test set	TSs predicted by HMMTOP	158	144 (89%)
	TSs taken from SwissProt annotations	158	149 (94%)
Möller *et al.*	Set A	24	22 (92%)
	Non-redundant set	121	107 (88%)

[a]The method fails on the proteins TRSR_HUMAN and CYOA_ECOLI (SwissProt codes) of the set of 72 proteins (Pasquier *et al.*, 1999).

The algorithm was also tested on a set of 158 transmembrane proteins used by Tusnády and Simon (Tusnády and Simon, 1998) to assess the accuracy of their HMM (Hidden Markov Model) algorithm (HMMTop) employed for topology prediction (Table I). This test set is a collection of three different data sets used earlier for transmembrane prediction methods: 83TMP (Jones *et al.*, 1994), 48TMP (Rost *et al.*, 1996) and prokTMP (Cserzö *et al.*, 1997). Taking into account the SwissProt annotation of transmembrane segments for this set, our tool predicts correctly the topology of 149 (94%) transmembrane proteins and falsely 9 (6%) of them (falsely predicted are the proteins with SwissProt codes: ALKB_PSEOL, ATPL_ECOLI, CYOA_ECOLI, GLR1_RAT*, GPT_CRILO*, RFBP_SALTY, SPG1_STRSP, SSRG_RAT*, TRSR_HUMAN*; marked with asterisks are eukaryotic proteins).

Assuming as transmembrane segments those predicted by HMMTop (Tusnády and Simon, 1998), our algorithm, OrienTM, predicts the correct topology for 141 proteins (89%), whereas HMMTop itself predicts correctly 143 (90%) with the same assumption. In this case, our method fails on the following 17 proteins: ATPL_ECOLI, BACH_HALSS, CYDA_ECOLI, CYDB_ECOLI, CYOD_ECOLI, GAC1_RAT*, GAC3_

MOUSE*, HISM_SALTY, KDPD_ECOLI, MAGL_MOUSE*, MTR_ECOLI, RFBP_SALTY, SECD_ECOLI, SPG1_STRSP, SSRG_RAT*, TOLQ_ECOLI, TRSR_HUMAN*. Those marked with asterisks are eukaryotic proteins.

The prediction efficiency of OrienTM was also tested on another set of integral membrane proteins presented recently (Möller *et al.*, 2000). This test set was created to contain a collection of transmembrane proteins with annotated transmembrane regions, for which good experimental evidence exists. It is available at the address ftp://ftp.ebi.ac.uk/databases/testsets/transmembrane. The intention of the authors was to present a reliable, perhaps universally accepted, set to benchmark the performance of transmembrane prediction programs. The authors subdivide this set into subsets of different trust levels A, B, C, D: the A subset contains transmembrane proteins whose structure is available (highest trust level), the B subset contains transmembrane proteins for which very good biochemical characterization with at least two complementary methods exists, the C subset has proteins for which basic biochemical characterization was done and the D subset contains transmembrane proteins for which no biochemical characterization is available. A non-redundant data set is also provided by the authors, containing non-homologous entries of trust levels A–C.

OrienTM was applied to test set A and also to the non-redundant data set (Table I). Set A contains 37 proteins. Excluding mitochondrial proteins and proteins of unknown topology, a set of 24 proteins was tested. The topology for 22 of them was predicted correctly (92%). The method fails to predict the topology of two prokaryotic proteins ATPL_ECOLI and KCSA_STRLI (SwissProt codes). The non-redundant data set contains 148 proteins. Similarly, excluding mitochondrial and of unknown topology proteins, this test set is reduced to 121 proteins. OrienTM predicts correctly the topology for 107 of them (88% accuracy) and fails on the following 14 proteins: ATPL_ECOLI, SPG1_STRSP, KCSA_STRLI, TRSR_HUMAN, PTND_ECOLI, EBR_STAAU, GEF_ECOLI, GSPL_PSEAE, GSPN_ERWCA, LEP3_ERWCA, FDOI_ECOLI, DCRA_DESVH, CAN1_YEAST, CVAA_ECOLI, TOLR_ECOLI (SwissProt codes). Interestingly, 12 of them are prokaryotic and only two (TRSR_HUMAN and CAN1_YEAST) eukaryotic. Again, it seems that the algorithm performs much better, with a high accuracy level, on eukaryotic transmembrane proteins. Obtaining parameters separately for eukaryotic and prokaryotic proteins does not improve the accuracy of the predictions. A possible cause of the reduced accuracy of prediction for prokaryotic proteins might be that initial annotations for this class of membrane proteins are more erroneous than for eukaryotic proteins.

Discussion

It appears that a simple statistical method such as OrienTM, which derives position-specific information for residues located at extra- and intracellular non-transmembrane segments performs at least as well as the most popular existing methods (von Heijne, 1992; Sipos and von Heijne, 1993; Jones *et al.*, 1994; Fariselli and Casadio, 1996; Rost *et al.*, 1996; Diederichs *et al.*, 1998; Tusnády and Simon, 1998).

Determination of transmembrane segment (TS) endpoints (particularly if this is carried out by computational methods) frequently is not absolutely accurate. That is, some TSs may be missed or over-predicted and also their exact endpoints may be wrongly annotated. It would seem dangerous, then, to use such data in our method, the statistics of which are position-specific. However, comparison between predictions for the same protein set as described in SwissProt and as predicted by prediction algorithms shows that predictions are still accurate. Even if one or sometimes two segments are missed and the rest of them are moved, the topology predicted by OrienTM is almost always close to the correct topology.

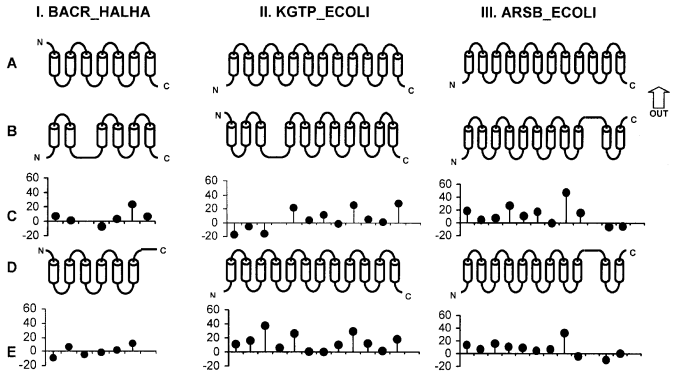

Fig. 1. Examples of use of OrienTM as a topology post-processor on three proteins with known topology supported with good experimental data: (I) bacteriorhodopsin (BACR_HALHA; Edman *et al*, 1999); (II) α-ketoglutarate permease (KGTP_ECOLI; Seol *et al*, 1993); and (III) arsenical translocation ATPase subunit B (ARSB_ECOLI; Wu *et al*, 1992). In all three cases, several known TS (transmembrane segment) prediction algorithms (data not shown), including our algorithm PRED-TMR (Pasquier *et al.*, 1999), miss one TS. The arrow indicates the direction from intra- to extracytoplasmic regions. For each protein: (**A**) experimentally determined topology. (**B**) Topology prediction based on TSs proposed by PRED-TMR (Pasquier *et al.*, 1999). PRED-TMR misses one TS in each protein. (**C**) A graph of the absolute values of individual TS scores versus TS relative number (counting from the N-terminus). Data points are taken as positive values if the corresponding scores agree with the predicted by OrienTM model, otherwise they are taken as negative. As seen, a sign inversion occurs, when a TS is missing. (**D**) Topology prediction based on TSs proposed by TMHMM (Krogh *et al.*, 2001). TMHMM predicts correctly all TSs of protein KGTP_ECOLI and misses one TS of proteins BACR_HALHA and ARSB_ECOLI. (**E**) As in (C) but with TSs predicted by TMHMM (Krogh *et al.*, 2001). OrienTM suggests that the topology is correct for KGTP_ECOLI (no sign inversion), indicates the position of the missing TS in ARSB_ECOLI (where the sign inversion occurs) and does not help in the suggestion of the correct topology for BACR_HALHA, where TMHMM misses the last TS.

The statistical parameters used to predict the topology can also easily be used in order to evaluate TS predictions, carried out by any experimental or computational tool. Nevertheless, the total score calculated for a topology scheme is useless if the TS determination is very erroneous. However, individual scores of each TS or NTS can help. Observing, for example, a cluster of TSs with scores opposite to the proposed topology, it is an indication that perhaps another missed TS exists right before them. Figure 1 shows the application of this technique to three proteins with topology supported from good experimental data: bacteriorhodopsin (BACR_HALHA; Edman *et al*, 1999), α-ketoglutarate permease (KGTP_ECOLI; Seol *et al*, 1993) and arsenical translocation ATPase subunit B (ARSB_ECOLI; Wu *et al*, 1992). In all three cases, several known TS prediction algorithms (data not shown), including our algorithm PRED-TMR, miss one TS. The position of the missed TS, in most cases, is suggested by the OrienTM scores (Figure 1, I–III). Therefore, it appears that OrienTM can be used as a topology post-processor, which also aims at enhancing the accuracy of topology predictions.

Availability

The program may be used freely through the Internet, at the address http://biophysics.biol.uoa.gr/OrienTM, utilizing any navigation tool (e.g. Netscape, Internet Explorer).

Acknowledgement

The authors acknowledge the support of the EEC–TMR 'GENEQUIZ' grant ERBFMRXCT960019.

References

Aloy,P., Cedano,J., Olivia,B., Aviles,X. and Querol,E. (1997) *Comput. Appl. Biosci.*, **13**, 231–234.
Bairoch,A. and Apweiler,R. (1998) *Nucleic Acids Res.*, **26**, 38–42.
Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) *Nucleic Acids Res.*, **28**, 235–242.
Cserzö,M., Wallin,E., Simon,I., von Heijne,G. and Elofsson,A. (1997) *Protein Eng.*, **10**, 673–676.
Diederichs,K., Freigang,G., Umhau,S., Zeth,K. and Breed,J. (1998) *Protein Sci.*, **7**, 2413–2420.
Edman,K., Nollert,P., Royant,A., Belrhali,H., Pebay-Peyroula,E., Hajdu,J., Neutze,R. and Landau,E.M. (1999) *Nature*, **401**, 822–826.
Fariselli,P. and Casadio,R. (1996) *Comput. Appl. Biosci.*, **12**, 41–48.
Jennings,M.L. (1989) *Annu. Rev. Biochem.*, **196**, 283–298.
Jones,D.T., Taylor,W.R. and Thornton,J.M. (1994) *Biochemistry*, **33**, 3038–3049.
Krogh,A., Larsson,B., von Heijne,G. and Sonnhammer,E.L.L. (2001) *J. Mol. Biol.*, **305**, 567–580.
Möller,S., Kriventseva,E.V. and Apweiler,R. (2000) *Bioinformatics*, **16**, 1159–1160.
Pasquier,C., Promponas,V.J., Palaios,G.A., Hamodrakas,J.S. and Hamodrakas,S.J. (1999) *Protein Eng.*, **12**, 381–385.
Persson,B. and Argos,P. (1994) *J. Mol. Biol.*, **237**, 182–192.
Promponas,V.J., Palaios,G.A., Pasquier,C.M., Hamodrakas,J.S. and Hamodrakas,S.J. (1999) *In Silico Biol.*, **3**, 1–4.
Rost,B., Fariselli,P. and Casadio,R. (1996) *Protein Sci.*, **5**, 1704–1718.
Seol,W. and Shatkin,A.J. (1993) *J. Bacteriol.*, **175**, 565–567.
Sipos,L. and von Heijne,G. (1993) *Eur. J. Biochem.*, **213**, 1333–1340.
Tusnády,G.E. and Simon,I. (1998) *J. Mol. Biol.*, **283**, 489–506.
von Heijne,G. (1992) *J. Mol. Biol.*, **225**, 487–494.
von Heijne,G. (1996) *Prog. Biophys. Mol. Biol.*, **66**, 113–139.
Wu,J., Tisa,L.S. and Rosen,B.P. (1992) *J. Biol. Chem.*, **267**, 12570–12576.

Received November 28, 2000; revised March 3, 2001; accepted March 23, 2001