Consensus predictions of membrane protein topology

Johan Nilsson[a,b], Bengt Persson[a,b], Gunnar von Heijne[a,*]

[a]*Stockholm Bioinformatics Center, Department of Biochemistry, Stockholm University, S-106 91 Stockholm, Sweden*
[b]*Department of Medical Biochemistry and Biophysics, Karolinska Institutet, S-171 77 Stockholm, Sweden*

Received 16 November 2000; accepted 22 November 2000

First published online 1 December 2000

Edited by Matti Saraste

Abstract We have explored the possibility that consensus predictions of membrane protein topology might provide a means to estimate the reliability of a predicted topology. Using five current topology prediction methods and a test set of 60 *Escherichia coli* inner membrane proteins with experimentally determined topologies, we find that prediction performance varies strongly with the number of methods that agree, and that the topology of nearly half of all *E. coli* inner membrane proteins can be predicted with high reliability (> 90% correct predictions) by a simple majority-vote approach. © 2000 Federation of European Biochemical Societies. Published by Elsevier Science B.V. All rights reserved.

Key words: Membrane protein; Topology; Prediction; Bioinformatics

1. Introduction

Computational methods for identifying potential integral membrane proteins and predicting their topology from their amino acid sequence have become increasingly important as a result of the genome sequencing projects. Current estimates put the fraction of integral membrane proteins in a typical genome between 20% and 25% [1], and even slight improvements in the ability to predict membrane protein topology will have major effects on, e.g. automatic sequence annotation efforts.

Here, we have explored a very simple way of estimating the reliability of a topology prediction by combining the results from five currently much used methods according to a 'majority-vote' principle. We show that the fraction of correctly predicted topologies over a test set of 60 *Escherichia coli* inner membrane proteins with experimentally determined topologies goes up with the number of methods that agree on the prediction, and is close to one when four or more methods agree. Four or five methods agree for 53% of the proteins in the test set, and for 46% of 764 proteins from *E. coli* that are identified as inner membrane proteins by the TMHMM method [1]. It thus appears that highly reliably topology predictions can be made for a substantial subset of all bacterial inner membrane proteins by the simple requirement that different prediction methods agree on the result.

2. Materials and methods

2.1. Test set proteins with experimentally determined topology

A test set was extracted from a recently assembled collection of membrane proteins with experimentally determined topologies [2] by including only *E. coli* proteins of 'trust levels' A–C, i.e. proteins for which reliable experimental topology information is available (type C proteins with partial topologies were excluded). *E. coli* proteins OPPB_ECOLI and OPPC_ECOLI were added since close homologs (identity > 95%) from *Salmonella typhimurium* were present in the collection. 12 additional proteins were collected by us from the recent literature (PNTA_ECOLI, PNTB_ECOLI [3], PUTP_ECOLI [4], DSBD_ECOLI [5], PROW_ECOLI [6], GABP_ECOLI [7], MDOH_ECOLI [8], YRBG_ECOLI (our unpublished data), YDGQ_ECOLI, ORF193 [9], NHAA_ECOLI [10], DCUA_ECOLI [11]). In total, the test set contained 60 proteins with the following SwissProt identifiers: AMTB_ECOLI, ARSB_ECOLI, ATP6_ECOLI, ATPL_ECOLI, CODB_ECOLI, CPXA_ECOLI, CYDA_ECOLI, CYDB_ECOLI, CYOB_ECOLI, CYOC_ECOLI, CYOD_ECOLI, CYOE_ECOLI, DCUA_ECOLI, DHG_ECOLI, DMSC_ECOLI, DSBB_ECOLI, DSBD_ECOLI, EXBB_ECOLI, FDOI_ECOLI, FRDC_ECOLI, FRDD_ECOLI, FTSH_ECOLI, GABP_ECOLI, HLYB_ECOLI, KDGL_ECOLI, KDPD_ECOLI, KGTP_ECOLI, KPM1_ECOLI, LACY_ECOLI, LEP_ECOLI, LSPA_ECOLI, LYSP_ECOLI, MALF_ECOLI, MALG_ECOLI, MDOH_ECOLI, MELB_ECOLI, MSCL_ECOLI, MTR_ECOLI, NHAA_ECOLI, OPPB_ECOLI, OPPC_ECOLI, PHEP_ECOLI, PNTA_ECOLI, PNTB_ECOLI, PROW_ECOLI, PTNC_ECOLI, PUTP_ECOLI, RBSC_ECOLI, RHAT_ECOLI, SECD_ECOLI, SECE_ECOLI, SECY_ECOLI, TCR1_ECOLI, TCR2_ECOLI, TOLQ_ECOLI, TRD1_ECOLI, UHPT_ECOLI, YDGQ_ECOLI, YRBG_ECOLI, ORF193.

2.2. Identification of E. coli inner membrane proteins

The full set of *E. coli* ORFs was downloaded from the EcoGene database [12] at http://genolist.pasteur.fr/Colibri/ and putative membrane proteins with a minimum of two predicted transmembrane helices were identified by TMHMM [1].

2.3. Prediction methods

Five topology prediction methods – TMHMM [1,13], HMMTOP [14], MEMSAT [15], TOPPRED [16,17], and PHD [18] – were used in their single-sequence mode (i.e. information from homologous proteins was not included). All user-adjustable parameters were left at their default values. For TOPPRED, if the overall bias in (Lys+Arg) residues was zero, the orientation of the protein was predicted based on the net charge difference across the most N-terminal transmembrane segment, or, if this was also zero, on the overall amino acid bias between the even- and odd-numbered loops [17]. Predictions were counted as correct if they matched the experimentally determined number of transmembrane helices and the location of the protein's N-terminus (cytoplasmic or periplasmic). Likewise, two predictions were considered to agree if the predicted number of transmembrane helices and the location of the N-terminus were the same. Thus, the exact beginning and end of each transmembrane helix in the sequence was not scored, as this information is available only for proteins with a known three-dimensional (3D) structure (trust level A).

*Corresponding author. Fax: (46)-8-15 36 79.
E-mail: gunnar@biokemi.su.se

0014-5793/00/$20.00 © 2000 Federation of European Biochemical Societies. Published by Elsevier Science B.V. All rights reserved.
PII: S0014-5793(00)02321-8

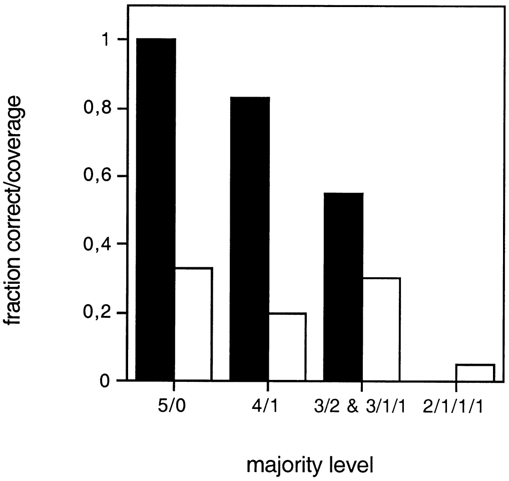

Fig. 1. Fraction of correctly predicted topologies (black bars) and fraction of the test set covered (white bars) for different levels of agreement between the five prediction methods (5/0, all methods agree; 4/1, four methods agree; 3/2, three methods agree, the remaining two agree with each other; 3/1/1, three methods agree, the remaining two do not agree with each other, etc.).

3. Results

In this study, we have used five popular topology prediction methods: TMHMM [1,13], HMMTOP [14], MEMSAT [15], TOPPRED [16,17], and PHD [18]. All five methods are designed to identify potential transmembrane α-helices and to predict the overall in–out topology of the protein in the membrane. TMHMM and HMMTOP both use a hidden Markov model formalism to describe the 'architecture' of an integral membrane protein, PHD is based on a neural network predictor, MEMSAT uses dynamic programming to optimally 'thread' a polypeptide chain through a set of topology models, and TOPPRED identifies 'certain' and 'putative' transmembrane α-helices from a standard hydrophobicity plot and then chooses the most likely topology based on the 'positive inside' rule [19]. All five methods use the known asymmetric distribution of amino acids between the *cis*- and *trans*-sides of the membrane in the prediction.

Since our immediate aim was to improve topology prediction for bacterial inner membrane proteins, we chose a test set of *E. coli* proteins with experimentally determined topologies from a carefully curated, recently collected database [2]. 12 additional proteins with known topologies were found in the recent literature (see Section 2). While some of these proteins have been used as training examples in the construction of the five methods, we felt that this independently collected set would nevertheless provide a good indication of prediction performance. We have refrained from including eukaryotic membrane proteins in this study, since (i) the experimental methods available for determining their topology are somewhat less reliable than those used for bacterial inner membrane proteins and often generate controversies [20–22], and (ii) they often have cleavable N-terminal signal peptides that complicate the prediction [1].

Table 1
Fraction of correctly predicted topologies over the test set of 60 proteins for the five methods used in this study

Method	Fraction correct predictions
TMHMM	0.72
HMMTOP	0.73
MEMSAT	0.67
TOPPRED	0.60
PHD	0.48

The fractions of correctly predicted topologies for the five methods taken individually are given in Table 1. They are slightly worse than the corresponding values reported in the original publications, but in general agree quite well with our expectations. It appears that the two most recent methods – TMHMM and HMMTOP – perform best and both make roughly 75% correct predictions.

We next classified all proteins in the test set according to the number of methods that gave the same predicted topology. As seen in Fig. 1, all five methods agree for 20 of the 60 proteins (a coverage of 33%), and the predicted topology is correct in all cases. Similarly, when four out of five methods agree, 10 out of 12 topologies predicted by the majority are correct. In the 18 cases where three methods agree, 10 predicted topologies are correct, and when only two methods agree, none of the three predicted topologies is correct. Thus, more than half of the proteins in the test set are predicted with a reliability better than 90% (the 32 cases where at least four methods agree).

There are fewer errors when the majority includes the two best-performing methods (TMHMM and HMMTOP; data not shown), although the numbers are too small to give reliable statistics when all different combinations of methods are compared.

Interestingly, there are seven proteins for which none of the methods predict the experimentally determined topology, even though as many as four out of the five methods agree in two cases, Table 2. We have scrutinized the experimental data for these two proteins, and find that they do not rule out the majority prediction for CYOE_ECOLI. Also, DCUA_ECOLI has a rather hydrophobic C-terminal tail that four of the five methods predict to span the membrane with the C-terminus in the cytoplasm. According to the high activity of a C-terminal

Table 2
Test set proteins for which none of the five methods predict the correct topology

SwissProt ID	Experimental	TMHMM	HMMTOP	MEMSAT	TOPPRED	PHD
HLYB_ECOLI	8-c	5-c	8-p	6-c	6-c	6-c
ARSB_ECOLI	12-c	11-c	14-p	13-c	13-p	13-p
RBSC_ECOLI	6-c	8-c	7-p	9-c	10-c	9-c
CYDA_ECOLI	7-c	9-p	9-p	1-c	9-p	8-c
CYOE_ECOLI	7-c	9-c	9-c	9-c	9-c	10-p
NHAA_ECOLI	12-c	11-c	10-c	11-c	11-p	10-c
DCUA_ECOLI	10-p	11-p	11-p	11-p	11-p	11-c

The experimentally determined as well as the individual topology predictions are given in the respective column as the number of transmembrane helices followed by the location of the N-terminus (c = cytoplasmic, p = periplasmic).

Table 3
Fractions of 764 predicted *E. coli* inner membrane proteins with different levels of majority predictions (5/0, all methods agree; 3/2, three methods agree, the remaining two agree with each other; 3/1/1, three methods agree, the remaining two do not agree with each other, etc.)

Majority level	Fraction of *E. coli* membrane proteins
5/0	0.22
4/1	0.24
3/2	0.10
3/1/1	0.17
2/1/1/1	0.13
No majority	0.14

β-lactamase fusion, the hydrophobic tail is periplasmic; however, it may be that the artificial lengthening of the C-terminus in this construct may alter the topology of the C-terminal tail, which would reconcile the predicted and observed topologies. It is thus possible that in both these cases the majority prediction is correct, and that there are no incorrect predictions also when four of the five methods agree. For the remaining five proteins, however, it appears that the majority predictions are incorrect.

Given the encouraging results on the test set, we also applied the consensus approach to 764 putative *E. coli* inner membrane proteins identified by TMHMM [1]. As shown in Table 3, the fraction of these proteins where four or five methods gave the same prediction was 46%, suggesting that a very reliable topology prediction can be made for nearly half of the *E. coli* inner membrane proteins. A list of those proteins and their predicted topologies can be found at http://www.sbc.su.se/∼johan/Very_Reliable_Topol_Pred.html.

In summary, the reliability of a topology prediction can be estimated by the number of prediction methods that agree: the larger the majority vote, the more likely is the prediction to be correct. By combining a number of prediction methods, proteins with a particularly clear-cut pattern of strongly hydrophobic transmembrane helices and a strong amino acid bias between the cytoplasmic and periplasmic loops – e.g. the 'easy-to-predict' proteins – can be identified. Whether the same holds true also for eukaryotic membrane proteins needs to be tested. We suggest that large-scale sequence annotation efforts may profitably use a battery of topology prediction methods to allow the user to get an idea of how much trust to place in a given prediction.

Acknowledgements: This work was supported by a center grant from the Foundation for Strategic Research to Stockholm Bioinformatics Center, a grant from the Swedish Technical Sciences Research Council to G.v.H., and grants from the Magnus Bergvall and Åke Wiberg Foundations to B.P.

References

[1] Krogh, A., Larsson, B., von Heijne, G. and Sonnhammer, E. (2000) J. Mol. Biol., in press.
[2] Möller, S., Kriventseva, E. and Apweiler, R. (2000) Bioinformatics, in press.
[3] Möller, J. and Rydström, J. (1999) J. Biol. Chem. 274, 19072–19080.
[4] Jung, H., Rubenhagen, R., Tebbe, S., Leifker, K., Tholema, N., Quick, M. and Schmid, R. (1998) J. Biol. Chem. 273, 26400–26407.
[5] Chung, J., Chen, T. and Missiakas, D. (2000) Mol. Microbiol. 35, 1099–1109.
[6] Haardt, M. and Bremer, E. (1996) J. Bacteriol. 178, 5370–5381.
[7] Hu, L.A. and King, S.C. (1998) Biochem. J. 336, 69–76.
[8] Debarbieux, L., Bohin, A. and Bohin, J.-P. (1997) J. Bacteriol. 179, 6692–6698.
[9] Sääf, A., Johansson, M., Wallin, E. and von Heijne, G. (1999) Proc. Natl. Acad. Sci. USA 96, 8540–8544.
[10] Rothman, A., Padan, E. and Schuldiner, S. (1996) J. Biol. Chem. 271, 32288–32292.
[11] Golby, P., Kelly, D., Guest, J. and Andrews, S. (1998) J. Bacteriol. 180, 4821–4827.
[12] Rudd, K. (2000) Nucleic Acids Res. 28, 60–64.
[13] Sonnhammer, E., von Heijne, G. and Krogh, A. (1998) Intell. Syst. Mol. Biol. 6, 175–182.
[14] Tusnady, G.E. and Simon, I. (1998) J. Mol. Biol. 283, 489–506.
[15] Jones, D.T., Taylor, W.R. and Thornton, J.M. (1994) Biochemistry 33, 3038–3049.
[16] von Heijne, G. (1992) J. Mol. Biol. 225, 487–494.
[17] Claros, M.G. and von Heijne, G. (1994) Comput. Appl. Biosci. 10, 685–686.
[18] Rost, B., Fariselli, P. and Casadio, R. (1996) Protein Sci. 5, 1704–1718.
[19] von Heijne, G. (1986) EMBO J. 5, 3021–3027.
[20] Contifine, B.M., Lei, S.J. and McLane, K.E. (1996) Annu. Rev. Biophys. Biomol. Struct. 25, 197–229.
[21] Jennings, M.L. (1989) Annu. Rev. Biochem. 58, 999–1027.
[22] van Geest, M. and Lolkema, J.S. (2000) Microbiol. Mol. Biol. Rev. 64, 13–33.