Bioinformatics, 32(8), 2016, 1158–1162
doi: 10.1093/bioinformatics/btv709
Advance Access Publication Date: 7 December 2015
Original Paper

OXFORD

Structural bioinformatics

Improved topology prediction using the terminal hydrophobic helices rule

Christoph Peters[1], Konstantinos D. Tsirigos[1], Nanjiang Shu[1,2] and Arne Elofsson[1,*]

[1]Department of Biochemistry and Biophysics, Science for Life Laboratory and [2]Sweden Bioinformatics Infrastructure for Life Sciences (BILS), Stockholm University, Solna 17121, Sweden

*To whom correspondence should be addressed.

Associate Editor: Anna Tramontano

Received on 16 July 2015; revised on 20 November 2015; accepted on 29 November 2015

Abstract

Motivation: The translocon recognizes sufficiently hydrophobic regions of a protein and inserts them into the membrane. Computational methods try to determine what hydrophobic regions are recognized by the translocon. Although these predictions are quite accurate, many methods still fail to distinguish marginally hydrophobic transmembrane (TM) helices and equally hydrophobic regions in soluble protein domains. *In vivo*, this problem is most likely avoided by targeting of the TM-proteins, so that non-TM proteins never see the translocon. Proteins are targeted to the translocon by an N-terminal signal peptide. The targeting is also aided by the fact that the N-terminal helix is more hydrophobic than other TM-helices. In addition, we also recently found that the C-terminal helix is more hydrophobic than central helices. This information has not been used in earlier topology predictors.

Results: Here, we use the fact that the N- and C-terminal helices are more hydrophobic to develop a new version of the first-principle-based topology predictor, SCAMPI. The new predictor has two main advantages; first, it can be used to efficiently separate membrane and non-membrane proteins directly without the use of an extra prefilter, and second it shows improved performance for predicting the topology of membrane proteins that contain large non-membrane domains.

Availability and implementation: The predictor, a web server and all datasets are available at http://scampi.bioinfo.se/.

Contact: arne@bioinfo.se

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The biogenesis of most α-helical transmembrane (TM) proteins follows a path where the synthesis is halted after a signal peptide is recognized by a signal recognition particle (Rapoport *et al.*, 2004). After the recognition, the ribosome is transported to the ER-, organelle- or plasma-membrane, where translation starts again with the emerging peptide chain threaded through the translocon channel, additional TM helices being recognized and, in some cases, the signal peptide is cleaved off. Although it is clear that, for some proteins, additional rearrangements occur (Kauko *et al.*, 2010), most membrane proteins appear to follow a folding procedure where the recognized TM helices come together to pack optimally (Rapoport *et al.*, 2004). This means that the location of membrane helices is primarily determined by the recognition of hydrophobic segments by the translocon.

The positive-inside rule (von Heijne and Gavel, 1988) governs the overall orientation of a protein, so that there is a global pattern of positively charged amino acids in the loops that occur on the periplasmic side of the membrane. The exact way in which the positive-inside rule acts is unclear, but even the presence or lack of positively charged residues close to the C-terminal can shift the orientation of the protein entirely (Seppälä *et al.*, 2010). Ultimately, the combination of hydrophobic segments recognized by the translocon and the localization of positive charges determines the topology of membrane proteins.

All membrane protein topology predictors take into account the biological mechanisms for membrane protein biogenesis, consciously or non-consciously. Many different computational methods, including Hidden Markov Models (HMMs) (Käll *et al.*, 2004, 2005; Krogh *et al.*, 2001; Tusnady and Simon, 2001; Viklund and Elofsson, 2004, 2008), Support vector machines (SVMs) (Nugent and Jones, 2009), Dynamic Bayesian Networks (Reynolds *et al.*, 2008) and Artificial Neural Networks (Jones, 2007a; Viklund *et al.*, 2008) have been used in topology prediction. A common feature in all these methods is that they identify hydrophobic TM helices and then combine these predictions with the positive-inside rule in order to determine the topology. The positive-inside rule is used to allow for less hydrophobic TM-segments to be recognized if this results in that more positive residues are found in periplasmic loops. Exactly how the balance between the hydrophobicity cut-off and the positive-inside rule is optimized varies between methods. In some methods, such as TOPPRED (Claros and von Heijne, 1994) the two parts are performed independently, while in most others the optimization is done simultaneously.

In contrast to most other topology predictors, the SCAMPI predictor performs topology prediction using no additional information apart from the hydrophobicity of the segments and the positive-inside rule (Bernsel *et al.*, 2008). SCAMPI only requires that two parameters, namely the hydrophobicity weight and the relative weight of the positive-inside rule, are optimized. The performance of SCAMPI is comparable with other state-of-the-art methods that utilize much more complicated schemes and many more parameters (Viklund and Elofsson, 2008).

One frequent problem when predicting the topology is that TM-helices are often erroneously predicted in non-TM domains. This problem becomes stronger in predictors that show the best overall performance (Tsirigos *et al.*, 2012). The reason is that the hydrophobicity of the most hydrophobic segments in non-TM domains may overlap with the hydrophobicity of some TM helices, see Figure 1. This makes it impossible to use a single strict cut-off to separate membrane and non-membrane regions, see the overlapp in Figure 1.

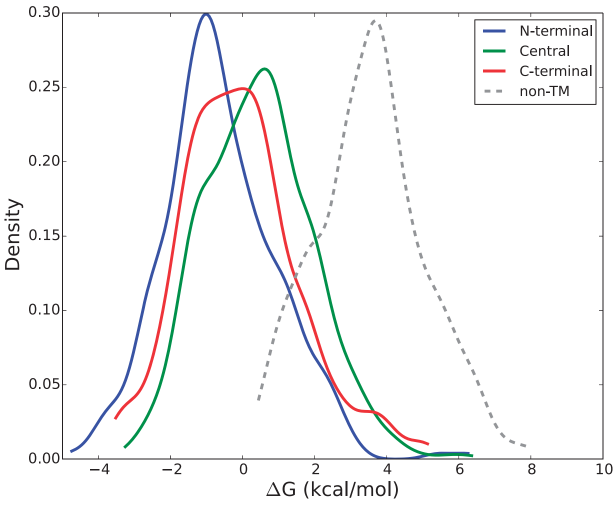

Fig. 1. Distributions of the hydrophobicity of the N-terminal, central and C-terminal helices are compared with the hydrophobicity of the most hydrophobic region located in a soluble domain. The overlap between the non-TM and TM hydrophobicity indicate regions that computationally are difficult to distinguish

Most non-TM proteins never see the translocon and therefore the insertion of hydrophobic non-TM segments is not a problem *in vivo* (Elofsson and von Heijne, 2007). Here, the targeting of proteins to the translocon is either guided by a signal peptide or by the most N-terminal TM-helix. In either case, the N-terminal helix is, on average, more hydrophobic than later helices (Hedin *et al.*, 2010), see Figure 1. From a prediction point-of-view, this means that, at least one helix in a TM-protein, has to be more hydrophobic than almost any segment in a non-TM protein. This fact has been used by many predictors by including a prefilter that determines whether a protein is a membrane or a non-membrane one (Bernsel *et al.*, 2008; Jones, 2007). In Figure 1, it can be seen that the overlap in hydrophobic regions of non-TM proteins (dotted line) is smaller for the N-terminal helix, than for the other TM-helices. Because of that, a stricter cut-off can be used in the prefilter, resulting in fewer falsely predicted TM-proteins. Although this strategy is useful for TM-protein identification, it does not solve the problem of falsely identifying TM helices in non-TM domains of a TM-protein.

In addition to the more hydrophobic N-terminal helix, we recently found that the C-terminal TM-helix is also more hydrophobic than central TM-helices (Virkki *et al.*, 2014), see Figure 1. To understand this, we have to take into account marginally hydrophobic helices, i.e. helices that under normal circumstances are not fully recognized by the translocon. These helices can be inserted if the number of positively charged residues in an inside loop is sufficient. However, we noted that the effect of the positive residues in the loop is weaker if they are located at the C-termini, i.e. if they are not followed by another TM-

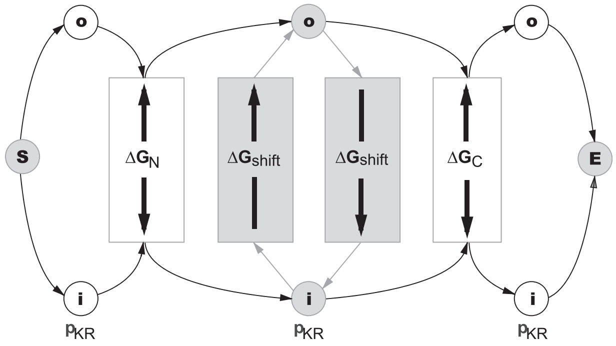

Fig. 2. A simplified version of the HMM used for SCAMPI2. Four parameters are used, the positive-inside bias P_{kr}, and three ΔG (biological hydrophobicity) cut-offs for N-terminal, central and C-terminal helices. These parameters are optimized individually. The arrows denote the orientation of the helices (from the cytoplasm to the extracellular space and *vice versa*). The grey parts show the original SCAMPI model

Table 1. The optimal parameters found during training of SCAMPI and SCAMPI2

Methods	p_{kr}	ΔG_N	ΔG_{shift}	ΔG_C	Filter
Methods using only single sequence					
SCAMPI	0.070	–	0.80	–	1.26
SCAMPI2	0.074	−1.60	1.00	−1.40	–
Methods using MSA					
SCAMPI	0.040	–	1.10	–	1.55
SCAMPI2	0.060	−1.80	1.30	−1.50	–

SCAMPI was trained both with and without the prefilter, i.e two or three parameters were optimized.

helix. It therefore seems as if the positive-inside rule is stronger when a following TM helix is present. Since no TM-helix follows the most C-terminal helix, the insertion of it cannot be aided by the positive-inside rule, i.e. it needs to be sufficiently hydrophobic to be inserted by itself (Virkki *et al.*, 2014).

To the best of the authors' knowledge, all topology predictors treat all TM-helices equally and do therefore ignore the fact that the N- and C-terminal TM-helices should be more hydrophobic. Here, we present a new version of SCAMPI that improves predictions by utilizing this observation.

2 Methods

In the original SCAMPI method, only two parameters, namely p_{KR} and ΔG_{shift}, are optimized, see Figure 2. p_{KR} stands for the positive-inside bias, which is the ratio of positively charged residues between the inside and the outside of the membrane. ΔG_{shift} is a hydrophobicity parameter used for the identification of TM helices. The topology is produced by finding the optimal path through an HMM with ΔG_{shift} and p_{KR} as the only optimizable parameters. However, for large-scale use, a strict prefilter based on the ΔG of the most hydrophobic region was employed in order to separate TM and non-TM proteins. In Table 1, it can be seen that both ΔG_{shift} and the prefilter cut-off are higher than 0, i.e. even helices that, experimentally, have <50% chance to be inserted into the membrane, have a good chance to pass the filter. The optimal ΔG_{shift} is actually close to the value that is best at separating TM and non-TM regions.

In the new version of SCAMPI, we introduce two additional parameters, ΔG_{N} and ΔG_{C}. These are the hydrophobicity parameters for the N- and C-terminal helices. For SCAMPI2, all four parameters were optimized based on a 3-fold cross-validation, using a grid-search (see Supplementary information for details). If several sets of parameters provide identical performance, they are all used. Therefore, there is an upper and lower limit in some of the numbers in Table 2. Similar to the original SCAMPI version, there are two flavors of SCAMPI2, one that uses single sequence and one that uses multiple sequence alignments (MSA). In the later case, the MSA is obtained using PSI-BLAST (Altschul *et al.*, 1997) with Uniref90 (Suzek *et al.*, 2015) as reference database. The MSA version is identical to the single sequence version but uses the average ΔG for all the sequences in the MSA, see Supplementary information.

TM proteins were extracted from the PDBTM database (Kozma *et al.*, 2013) and mapped to the respective Uniprot (UniProt Consortium, 2014) sequences using the SIFTS database (Velankar *et al.*, 2013). In contrast to some earlier studies, the full-length Uniprot sequences are used and not only the structurally determined regions. We assumed that no additional TM regions were present. A correct predicted topology requires that (i) the number of TM regions is correctly predicted, (ii) the location of each predicted helix overlaps with at least five residues with the observed helix and (iii) the N- and C-termini are located to be at the correct side of the membrane.

For topology assignment, we used PDBTM, OPM (Lomize *et al.*, 2006), TOPDB (Dobson *et al.*, 2014) and Uniprot annotations. After performing a 20% sequence identity homology reduction, 285 TM-proteins remained. Of them, 60 proteins contain non-membrane domains longer than 200 residues (denoted as long domains in Table 2). As a non-TM set 3597 non-TM proteins originating from SIGNALP4.0 (Petersen *et al.*, 2011) and previously used in TOPCONS (Tsirigos *et al.*, 2015) were used for evaluation. For the

Table 2. Fraction of proteins with correctly predicted topologies, for different versions of SCAMPI

Methods	Overall (%)	TM (%)	Long (%)	Non-TM (%)
Methods using only single sequence				
SCAMPI2	79–81	60–63	53–57	97–98
SCAMPI (Bernsel *et al.*, 2008)	67	61	35	72
SCAMPI-prefilter* (Bernsel *et al.*, 2008)	78	61	35	95
HMMTOP (Tusnady and Simon, 2001)	70	63	48	77
Philius† (Reynolds *et al.*, 2008)	81	67	50	94
Phobius† (Käll *et al.*, 2004)	76	56	63	95
TMHMM (Krogh *et al.*, 2001)	78	56	50	99
TOPCONS-single (Hennerdal and Elofsson, 2011)	82	69	60	95
Methods using MSA				
SCAMPI2	86	73	67	98-99
SCAMPI (Bernsel *et al.*, 2008)	74	72	43	75
SCAMPI + prefilter* (Bernsel *et al.*, 2008)	85	71	43	98
MEMSAT3* (Jones, 2007)	87	77	63	97
MEMSAT-SVM*,† (Nugent and Jones, 2009)	81	72	67	89
OCTOPUS (Viklund and Elofsson, 2008)	79	65	42	92
PolyPhobius† (Käll *et al.*, 2005)	81	67	65	95
PRO (Viklund and Elofsson, 2004)	80	62	43	97
PRODIV (Viklund and Elofsson, 2004)	38	66	47	13
SPOCTOPUS† (Viklund *et al.*, 2008)	75	72	63	78
TOPCONS (Tsirigos *et al.*, 2015)	88	78	70	97

Methods that are marked with an asterisk (*) include a prefilter, methods with dagger ($_{\dagger}$) also predict signal peptides. The accuracies are measured on the entire dataset (Overall), only the membrane proteins (TM), only the membrane proteins with long non-TM domains (Long) and the non-TM proteins (non-TM). The results for SCAMPI2 are cross validated, while the others are not. However, during the cross validation it is possible that in the training set several different parameters provide identical results. In such cases, all these parameters were tested and the maximum/minimum results are shown in the table. For TM proteins without a signal peptide, a topology is deemed correct when the predicted topology has the same N-terminal, same number of TM helices as the experimentally verified one and the helices overlap by at least five residues. For TM proteins with a signal peptide, a correct topology for SCAMPI (and other methods not designed for signal peptide prediction) is considered if no TM regions are predicted in the N-terminal part of the protein. The non-TM accuracy is given for the large dataset containing 3597 non-TM proteins and it requires that neither TM helix nor a signal peptide is predicted in these proteins cross-validation training, a randomly selected set of 285 non-TM proteins was used.

3 Results and discussion

When optimizing the parameters for SCAMPI2, $\Delta G_{\mathrm{Shift}}$ (describing the hydrophobicity cut-off for central helices) is set to 1.00 kcal/mol, slightly higher than the numbers in the original SCAMPI version (0.80), see Table 1. This provides the opportunity for less hydrophobic helices to be inserted, as opposed to the original SCAMPI. However, the cut-offs for N- and C-terminal helices, ΔG_{N} and ΔG_{C}, are set to be much stricter, -1.60 and -1.40, respectively. This results in that marginally hydrophobic regions that may occur before or after the TM regions are not erroneously predicted. This increases the accuracy for membrane proteins that contain non-TM domains longer than 200 residues, see Table 2. In Figure 3, we show four examples where SCAMPI2 avoids erroneous predictions.

The original single-sequence-based SCAMPI method predicts the correct topology for 61% of the proteins but only for 35% of the proteins with long non-TM domains, see Table 2. Furthermore, around one-third of non-TM proteins are wrongly classified as TM-proteins. This gives an overall accuracy in our benchmark of 67%. Adding a prefilter increases the overall accuracy to 78% by improving the separation between TM and non-TM proteins drastically. Using MSA, the overall accuracy increases to 74% without the prefilter and to 85% with the prefilter, but still only 43% of the proteins with long non-TM domains are predicted correctly.

The overall performance of SCAMPI2 is 79–81% when using single sequences and 86% when using MSA, see Table 2. More importantly, the prediction accuracy for the subset of proteins with the long non-TM domains is increased to 53–57 and 67%, respectively. Further, the discrimination between TM and non-TM proteins is slightly better in SCAMPI2 than when using the prefilter and SCAMPI. Still, ~40% (25% for MSA) of the topologies are wrongly predicted. An analysis of the errors indicates that 10% (3%) of the proteins are predicted in a wrong orientation of the entire protein in SCAMPI, while 25% (15%) are due to over or under-prediction of TM-helices and the rest because of misplaced helices.

In contrast to other state-of-the-art methods, SCAMPI2 only uses four parameters, yet its performance is comparable to more sophisticated predictors, see Table 2. For single-sequence methods, Philius shows a slightly better performance within the TM proteins, but Philius is not as accurate as SCAMPI2 for the separation of TM and non-TM proteins, which results in a similar overall performance. One important difference between Philius and SCAMPI2 is that Philius treats signal peptides separately, indicating a possible path toward further improvement of SCAMPI.

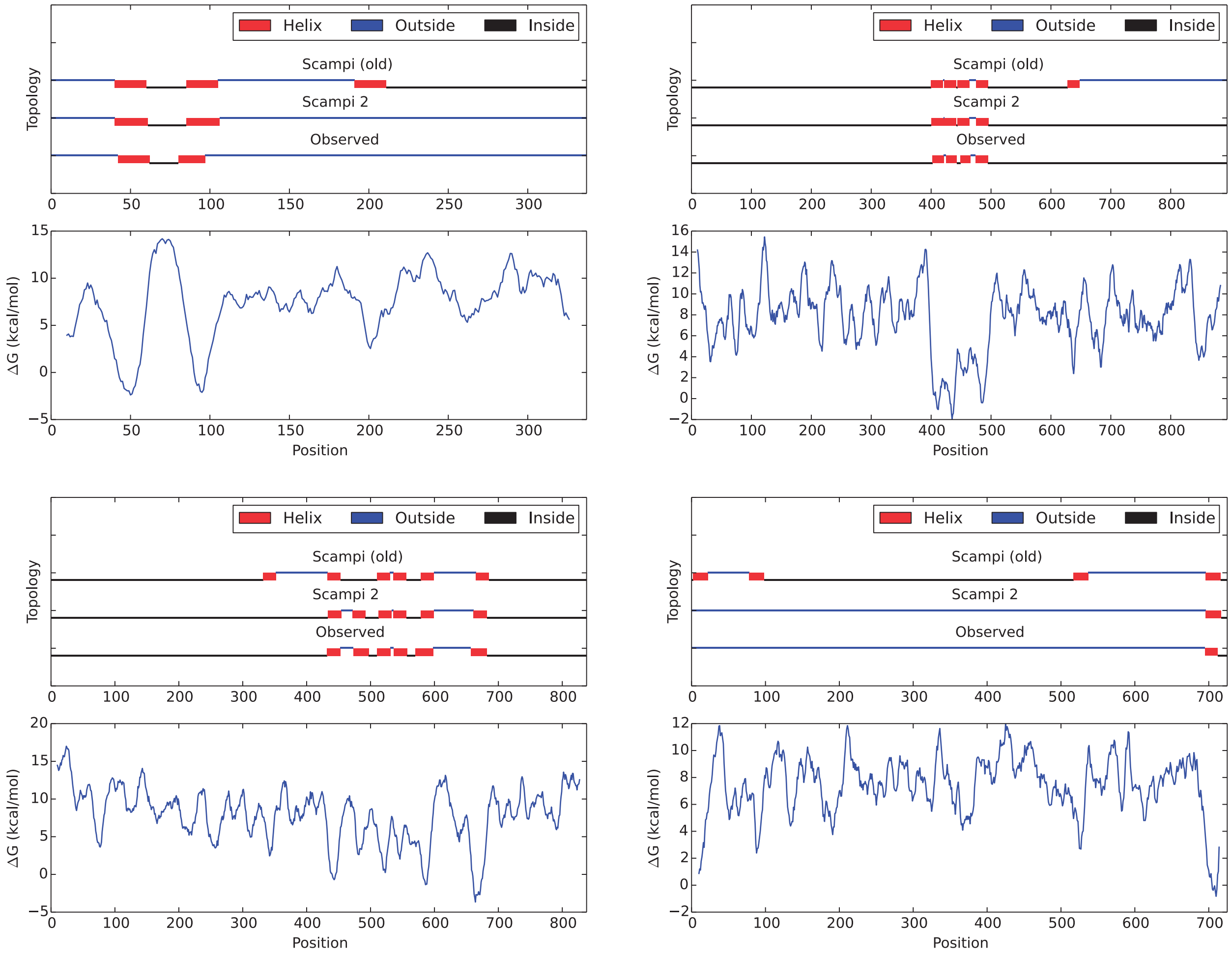

Fig. 3. Example of the improvement due to the new SCAMPI model for four proteins. (**a**) caa3-type cytochrome oxidase, chain B, (**b**) *Escherichia coli* histidine kinase receptor KdpD, chain A, (**c**) TRPV1 ion channel, chain G, (**d**) Cellulose synthase, chain B. By having a stricter cut-off for the N-terminal helix, the erroneously predicted helix is removed. The looser cut-off for central TM-helices and the different orientation of the preceding loops, aided the correct identification of the second TM-helix

For MSA methods, there is a similar picture, with MEMSAT3 showing a better topology predictions of TM proteins, but being slightly worse at separating TM and non-TM proteins. Finally, the consensus predictor TOPCONS that combines the predictions from several methods, including SCAMPI, shows a slightly higher performance than SCAMPI2.

However, SCAMPI2 has some additional advantages over the other methods; first it is significantly faster than Philius or MEMSAT3 (in single- or multiple-sequence modes, respectively). Moreover, the results presented here are cross validated, whereas for the rest of the methods there is certainly some overlap between our datasets and their training set.

Finally, we found it interestingly that the method that is best at separating TM and non-TM proteins is the 15-years-old TMHMM, with 99% accuracy. However, the good separation comes at the price of non-optimal accurate topology predictions. The accuracy is 56 versus 63% for HMMTOP developed at the same time. This highlights the importance and problems with recognition of marginally hydrophobic helices. If they are recognized, the topology predictions are better but there is a higher chance of false predictions in non-TM proteins.

4 Conclusions

We present a novel topology prediction method, SCAMPI2, that, to the best of the authors' knowledge, for the first time uses the observation that the N- and C-terminal helices in a TM-protein are, on average, more hydrophobic than the other TM-helices. By increasing the hydrophobicity requirement for the N- and C-terminal helices in TM proteins, it is possible to improve topology predictions, in particular for proteins with long non-TM domains. An additional advantage of this method is that no prefilter to distinguish TM and non-TM protein is required. Clearly, the difference in hydrophobicity of central and terminal helices could also be included in other methods, demonstrating therefore a possible improvement for all topology predictors, in particular if the hydrophobicity of signal peptides is also taken into account.

Funding

Funding was provided by the Swedish Research Council (VR-NT 2012-5046). N.S. was financed by Bioinformatics Infrastructure for Life Science (BILS).

Conflict of interest: none declared.

References

Altschul,S. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Bernsel,A. *et al.* (2008) Prediction of membrane-protein topology from first principles. *Proc. Natl Acad. Sci. USA*, **105**, 7177–7181.

Claros,M. and von Heijne,G. (1994) Toppred II: an improved software for membrane protein structure prediction. *Comput. Appl. Biosci.*, **10**, 685–686.

Dobson,L. *et al.* (2014) Expediting topology data gathering for the TOPDB database. *Nucleic Acids Res.*, **43**, D283–D289.

Elofsson,A. and von Heijne,G. (2007) Membrane protein structure: prediction versus reality. *Annu. Rev. Biochem.*, **76**, 125–140.

Hedin,L. *et al.* (2010) Membrane insertion of marginally hydrophobic transmembrane helices depends on sequence context. *J. Mol. Biol.*, **396**, 221–229.

Hennerdal,A. and Elofsson,A. (2011) Rapid membrane protein topology prediction. *Bioinformatics*, **27**, 1322–1323.

Jones,D. (2007) Improving the accuracy of transmembrane protein topology prediction using evolutionary information. *Bioinformatics*, **23**, 538–544.

Käll,L. *et al.* (2004) A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.*, **338**, 1027–1036.

Käll,L. *et al.* (2005) An HMM posterior decoder for sequence feature prediction that includes homology information. *Bioinformatics*, **21**, i251–i257.

Kauko,A. *et al.* (2010) Repositioning of transmembrane alpha-helices during membrane protein folding. *J. Mol. Biol.*, **397**, 190–201.

Kozma,D. *et al.* (2013) PDBTM: protein data bank of transmembrane proteins after 8 years. *Nucleic Acids Res.*, **41**, D524–D529.

Krogh,A. *et al.* (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.

Lomize,M. *et al.* (2006) OPM: orientations of proteins in membranes database. *Bioinformatics*, **22**, 623–625.

Nugent,T. and Jones,D. (2009) Transmembrane protein topology prediction using support vector machines. *BMC Bioinformatics*, **10**, 159.

Petersen,T. *et al.* (2011) Signalp 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods*, **8**, 785–786.

Rapoport,T. *et al.* (2004) Membrane-protein integration and the role of the translocation channel. *Trends Cell. Biol.*, **14**, 568–575.

Reynolds,S. *et al.* (2008) Transmembrane topology and signal peptide prediction using dynamic Bayesian networks. *PLoS Comput. Biol.*, **4**, e1000213.

Seppälä,S. *et al.* (2010) Control of membrane protein topology by a single c-terminal residue. *Science*, **328**, 1698–1700.

Suzek,B. *et al.* (2015) Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, **31**, 926–932.

Tsirigos,K. *et al.* (2012) A guideline to proteome-wide alpha-helical membrane protein topology predictions. *Proteomics*, **12**, 2282–2294.

Tsirigos,K. *et al.* (2015) The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. *Nucleic Acids Res.*, **43**, W401–W407.

Tusnady,G. and Simon,I. (2001) The HMMTOP transmembrane topology prediction server. *Bioinformatics*, **17**, 849–850.

UniProt Consortium (2014) Activities at the universal protein resource (uniprot). *Nucleic Acids Res.*, **42**, D191–D198.

Velankar,S. *et al.* (2013) SIFTS: structure integration with function, taxonomy and sequences resource. *Nucleic Acids Res.*, **41**, D483–D489.

Viklund,H. and Elofsson,A. (2004) Best alpha-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information. *Protein Sci.*, **13**, 1908–1917.

Viklund,H. and Elofsson,A. (2008) OCTOPUS: improving topology prediction by two-track ANN-based preference scores and an extended topological grammar. *Bioinformatics*, **24**, 1662–1668.

Viklund,H. *et al.* (2008) SPOCTOPUS: a combined predictor of signal peptides and membrane protein topology. *Bioinformatics*, **24**, 2928–2929.

Virkki,M. *et al.* (2014) The positive inside rule is stronger when followed by a transmembrane helix. *J. Mol. Biol.*, **426**, 2982–2991.

von Heijne,G. and Gavel,Y. (1988) Topogenic signals in integral membrane proteins. *Eur. J. Biochem.*, **174**, 671–678.