doi:10.1006/jmbi.2000.4315 available online at http://www.idealibrary.com on IDEAL® *J. Mol. Biol.* (2001) **305**, 567–580

JMB

Predicting Transmembrane Protein Topology with a Hidden Markov Model: Application to Complete Genomes

Anders Krogh[1]*, Björn Larsson[1], Gunnar von Heijne[2] and Erik L. L. Sonnhammer[3]

[1]*Center for Biological Sequence Analysis, Technical University of Denmark, Building 208 2800 Lyngby, Denmark*

[2]*Stockholm Bioinformatics Center, Department of Biochemistry, Stockholm University, S-106 91 Stockholm, Sweden*

[3]*Center for Genomics Research Karolinska Institutet, 171 77 Stockholm, Sweden*

**Corresponding author*

We describe and validate a new membrane protein topology prediction method, TMHMM, based on a hidden Markov model. We present a detailed analysis of TMHMM's performance, and show that it correctly predicts 97-98 % of the transmembrane helices. Additionally, TMHMM can discriminate between soluble and membrane proteins with both specificity and sensitivity better than 99 %, although the accuracy drops when signal peptides are present. This high degree of accuracy allowed us to predict reliably integral membrane proteins in a large collection of genomes. Based on these predictions, we estimate that 20-30 % of all genes in most genomes encode membrane proteins, which is in agreement with previous estimates. We further discovered that proteins with N_{in}-C_{in} topologies are strongly preferred in all examined organisms, except *Caenorhabditis elegans*, where the large number of 7TM receptors increases the counts for N_{out}-C_{in} topologies. We discuss the possible relevance of this finding for our understanding of membrane protein assembly mechanisms. A TMHMM prediction service is available at http://www.cbs.dtu.dk/services/TMHMM/.

Keywords: transmembrane helices; hidden Markov model; prediction of membrane protein topology; membrane proteins in genomes; protein structure prediction

Introduction

The prediction of transmembrane helices in integral membrane proteins is an important aspect of bioinformatics. The most successful methods to date not only predict individual transmembrane helices, but rather attempt to predict the full topology of the protein, i.e. the total number of transmembrane helices and their in/out orientation relative to the membrane (von Heijne, 1999). Reliable methods for discrimination between membrane proteins and soluble proteins and for topology prediction have important applications in genome analysis, and can be used to extract global trends in membrane protein evolution (Wallin & von Heijne, 1998).

Early methods for prediction of transmembrane helices used hydrophobicity analysis alone (see e.g. Argos *et al.*, 1982). Indeed some helices can be located with high reliability from a hydrophobicity plot, but others cannot. Another signal shown to be associated with transmembrane helices is the abundance of positively charged residues in the part of the sequence on the cytoplasmic side of the membrane, ''the positive inside rule'' (von Heijne, 1986, 1994). By combining charge bias analysis with hydrophobicity analysis, better predictions can be obtained (von Heijne, 1992). Although they vary in detail, almost all recent methods for prediction of transmembrane helices rely on those two signals. Several methods use a sliding window which is predicted as being part of a membrane helix or not, either by a weight matrix (Edelman, 1993) or by a neural network (Rost *et al.*, 1995; Casadio *et al.*, 1996). Some methods use multiple

Present address: B. Lärsson, Department of Quantum Chemistry, AIM Research School, Box 518, SE-75120 Uppsala, Sweden.

Abbreviations used: N_{in}, N terminus inside; N_{out}, N terminus outside; TM, transmembrane; HMM, hidden Markov model.

E-mail address of the corresponding author: krogh@cbs.dtu.dk

0022-2836/01/030567–14 $35.00/0

alignments to improve on the predictions (Persson & Argos, 1994; Rost *et al*., 1996).

Helical membrane proteins follow a ''grammar'' in which cytoplamic and non-cytoplasmic loops have to alternate. The grammar constrains the possible topologies, and thereby the possible transmembrane helices. Therefore, an integrated prediction method that takes the grammar into account can, in principle, give better results, even at the level of single transmembrane helix predictions. Jones *et al*. (1994) describe a dynamic programming algorithm that maximizes the total sum of residue scores, while at the same time obeying the grammar. Different scores were used for residues in the middle of a helix, in helix caps, and in loop regions. This method implicitly combines the hydrophobic signal and the charge bias signal into one integrated algorithm in a natural way.

Here we describe a new method, TMHMM, based on a hidden Markov model (HMM) approach (a preliminary description of TMHMM has been published by Sonnhammer *et al*., 1998). It resembles the method described by Jones *et al*. (1994), in that it has specialized modeling of various regions of a membrane protein: helix caps, middle of helix, regions close to the membrane, and globular domains. One of the main advantages of an HMM is that it is possible to model helix length, which has only been done fairly crudely in most other methods, by setting upper and lower limits for the length of a membrane helix. The HMM is very well suited for prediction of transmembrane helices because it can incorporate hydrophobicity, charge bias, helix lengths, and grammatical constraints into one model for which algorithms for parameter estimation and prediction already exist (see e.g. Durbin *et al*. (1998)). We further apply TMHMM to predict all membrane proteins in a large collection of mostly fully sequenced genomes, and present statistics on the frequency of proteins with different topologies. Interestingly, we find that proteins with both the N and C terminus in the cytoplasm dominate in almost all organisms.

Another HMM method, HMMTOP, has been independently developed (Tusnady & Simon, 1998). It builds on a very similar HMM architecture, but the method used for prediction is different. A model regularizer is estimated from a set of known transmembrane proteins, and for prediction a model is estimated from the query sequence and then used for predicting the structure of that sequence. The reported single-sequence prediction accuracy of HMMTOP, 78 % correct topology, is roughly the same as that of TMHMM, although comparing accuracies is difficult due to differences in datasets and cross-validation methods as discussed below.

Results

The TMHMM architecture

The layout of the model is shown in Figure 1(a). Each box in the drawing corresponds to a submodel designed to model a specific region of a membrane protein. These submodels contain several HMM states in order to model the lengths of the various regions. The arrows show how transitions between submodels can be made such that they obey the grammatical structure of the helical transmembrane proteins.

We have made no attempt to construct a sophisticated model of the globular domains of the transmembrane proteins, so the submodels labeled ''globular'' in Figure 1(a) are identical and consist of just one state with a transition to itself and to a loop state (see also Figure 1(b)). To capture the topogenic signal of the proteins, we model the residues close to the membrane in the submodels labeled ''loop'' and ''cap'', which are shown in Figure 1(b). Loops of lengths up to 20 residues are modeled by the loop model, whereas longer loops have to use the globular state. The transition topology ensures that any loop of length one or more is allowed. The 20 loop states of a loop submodel all have the same distribution of amino acid residues, but the three loop models are different. The cap submodels simply model the five first or last residues of the transmembrane regions. The model for the core of the transmembrane helices is shown in Figure 1(c). It is an array of 25 identical states with the possibility of jumping from one of the states (state 3 in the drawing) to many of the states down-stream. This topology models sequences of lengths between five and 25, which translates to helix lengths between 15 and 35 when the caps are included. This is consistent with the distribution of helix lengths in membrane proteins of known structure (Bowie, 1997). In this interval, the length distribution is explicitly represented by the transition probabilities of the transitions in the helix model. The HMM parameters, which are the probabilities of the 20 amino acid residues in the states and the probabilities that determine the length distributions of transmembrane helices etc., are estimated from a set of 160 proteins in which the locations of the transmembrane helices are known. The boundaries of the helices are often inaccurately determined (even from crystal structures), so we have designed an estimation procedure in which a model is used to redefine the boundaries.

Prediction of the transmembrane helices is done by finding the most probable topology given the HMM. This will give a set of exact helix boundaries. However, there are many almost equally probable ways to place the helix boundaries, and there are sometimes regions in the sequence that show weak signs of being transmembrane helices or predicted helices that have a fairly low probability. Such information is not contained in the most probable prediction. Therefore, we found it

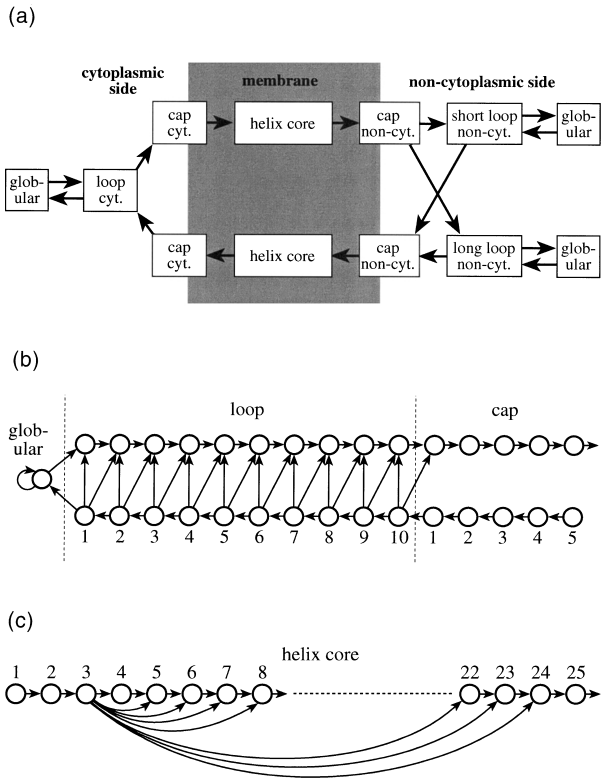

Figure 1. The layout of the hidden Markov model. (a) The overall layout. Each box corresponds to one or more states in the HMM. Parts of the model with the same text are tied, i.e. their parameters are the same. Cyt. represents the cytoplasmic side of the membrane and non-cyt. the other side. (b) The detailed structure of the inside and outside loop models and helix cap models. (c) The structure of the model for the helix core modelling lengths between five and 25, which translates to helices between 15 and 35 when the caps are included.

very useful also to use the three probabilities that a given residue is in a transmembrane helix, is on the cytoplasmic side, or on the periplasmic side. This additional information, which can be shown graphically as in Figure 2, shows where the prediction is certain and what alternatives there might be.

Analysis of correct and incorrect predictions

There are several types of mis-prediction that can occur when predicting the topology of a membrane protein. The simplest errors are over-predictions and under-predictions, i.e. predicting a transmembrane region where none is present or missing a true transmembrane region. Another type of error is that two adjoining transmembrane regions are joined together, so that they are predicted as a single long region, which we will term as a ''false merge''. Similarly, a long transmembrane region can be falsely predicted as being two short regions, here termed a ''false split''. Of course, all the helices can be predicted correctly, but the overall topology can be predicted as the inverse of the real topology, i.e. an inverted topology. A predicted transmembrane helix is considered correct if it overlaps by at least five residues with a real helix. If this fails, it is considered a shifted prediction if there is an overlap of at least one amino acid with the real helix. Known signal peptides were not removed from the sequences and were not counted as valid transmembrane helices. Signal peptides are sometimes hard to distinguish from transmembrane helices, as discussed below.

Table 1 shows the occurrence of errors in a cross-validation experiment. The ten cross-validation models corresponding to the first column of numbers in Table 1 were used in the discrimination analysis below. Since the training algorithm has a stochastic element (see Materials and Methods), the accuracy can vary. The cross-validation experiment was therefore repeated 40 times, and averages and standard deviations were calculated. These data are also shown in Table 1.

Discrimination between non-membrane and membrane proteins

Although the method has been developed and optimized for correct prediction of the topology, it can also be used for discrimination between helical membrane proteins and other proteins. This can be achieved in several ways, and we have investigated discrimination based on the following values:

(1) The number of predicted transmembrane helices (abbreviated ''pred. no. TMH'').
(2) The expected number of residues in transmembrane helices (abbreviated ''exp. no. AA'').
(3) The expected number of transmembrane helices.

The first is simply a count of the number of helices in the most likely structure found by the model. If the expected number of residues in transmembrane helices is high, the probability that it is a helical membrane protein is also high. A threshold value can be determined from the data and used for discrimination. The shortest transmembrane helices are around 18 residues long, so the cut-off value should be close to that. If the expected number of transmembrane helices is around one or larger, it is likely to be a helical transmembrane protein. Since it is an expectation value, it is not an integer number, and again a threshold value can be determined.

These measures are calculated using the cross-validation models, i.e. for a membrane protein, the model is used which did not have the protein in the training set. For the non-membrane proteins, the averages over the ten cross-validation models were calculated. The three discrimination measures are of course correlated. Figure 3 shows the correlation between the predicted number of helices and

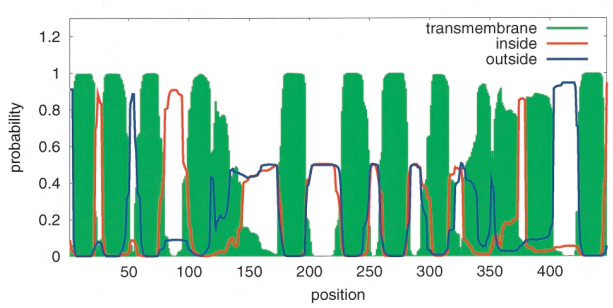

Figure 2. Posterior probabilities for a single sequence. The posterior probability for transmembrane helix, inside, or outside displayed for the gluconate permease 3 from *E. coli* (SWISS-PROT entry GNTP_ECOLI), for which the structure is unknown. Some parts of the protein are relatively certain, whereas other parts are less certain. It is unclear, for instance whether there are one or two transmembrane segments between amino acid 100 and 150, and between 325 and 375. This uncertainty is also reflected in a total uncertainty in which side the loops are (inside or outside) between 150 and 325. For this protein the single most probable topology turns out to have two helices in both of these regions giving 13 transmembrane helices in total, and this prediction turns out to be essentially identical to the annotation in SWISS-PROT. However, the posterior probability plot shows that the topology with only one helix in these regions (11 in total) is a quite likely alternative, whereas a topology with 12 or 14 transmembrane helices is not so likely because it would fit badly with the posterior probabilities of inside/outside in the two ends of the protein. In Klemm *et al.* (1996) 14 transmembrane helices are predicted for this protein; three helices are predicted in the region beween 100 and 150.

the expected number of transmembrane helices. These numbers are also correlated to the expected number of residues in transmembrane helices, as seen in Figure 4.

With all three measures, it is possible to identify all but one transmembrane protein in this data set with a very small number of false positives: for the rest of this work, we have used the expected number of residues in transmembrane helices. Figure 5 shows the discriminative power as a function of the cut-off used. At a cut-off of 18, which we have used below, the fraction of false positives is 0.5-1 % and around 1 % false negatives. The five proteins that are incorrectly classified as transmembrane in the cross-validation test are shown in Table 2. The chlorophyll *a-b* binding protein ab96 (Swissprot entry CB21_PEA) is the only membrane protein in the set of 160 that is classified as a non-membrane protein. This protein may be difficult to classify correctly because it is inserted into the thylakoid membrane.

For comparison, we have tested the ''maxH'' method (Boyd *et al.*, 1998) on the same data sets. It had seven false negatives (TMHMM had one) and three false positives (TMHMM had five). These numbers are valid for all *p*-value cut-offs between 0.04 and 0.68 for maxH. This test used the standard maxH program, so no cross-validation was performed, which might bias the result in favor of maxH.

Table 1. Types of errors

	Cross-validation		Mean and std. dev.	
Number of proteins	160			
of which single-spanning:	52	32.50 %		
Correctly predicted topology:	124	77.50 %	120.2	1.3
Invertedly predicted topology:	11	6.88 %	10.5	0.9
Correctly predicted N-terminal:	141	88.12 %	138.0	1.3
Under-predictions:	16	10.00 %	18.4	1.4
of which single-spanning:	1	0.62 %	0.6	0.5
Over-predictions:	12	7.50 %	14.1	0.6
of which single-spanning:	7	4.38 %	7.0	0.2
Both over- and under-predictions:	3	1.88 %	3.60	0.8
of which single-spanning:	1	0.62 %	0.58	0.5
Total number of real helices:	696			
Number of over-predicted helices:	17	2.44 %	20.1	0.6
Number of under-predicted helices:	19	2.73 %	21.7	1.8
Number of shifted helix predictions:	0		0.33	0.5
Number of falsely merged helices:	0		0.50	0.6
Number of falsely split helices:	0		0	0

The number of different types of errors in a cross-validated test of TMHMM. First column shows the cross-validation that is the basis for the discrimination analysis and the second column shows the average and standard deviation for 40 independent cross-validation experiments.

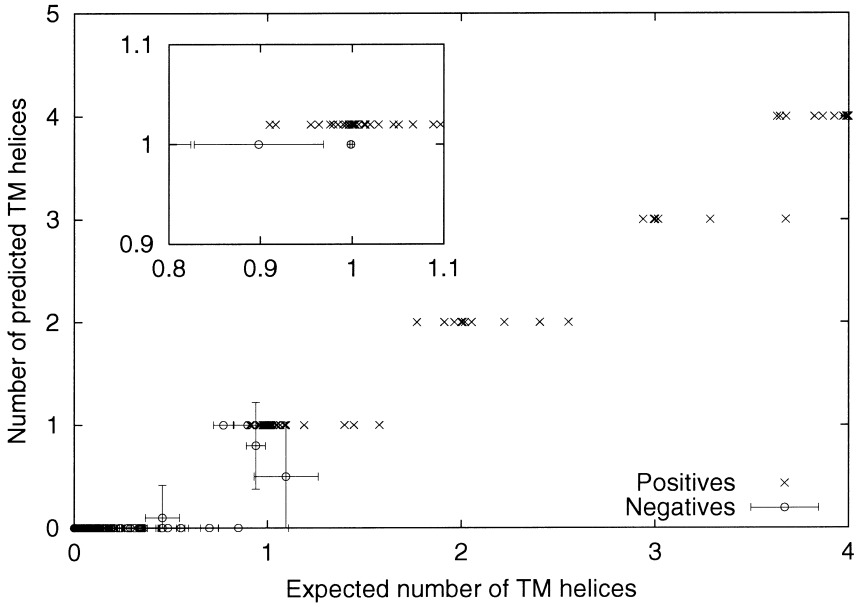

Figure 3. Correlation between the expected number of transmembrane helices and the predicted number of helices. The insert shows a blow-up of the critical region around 0.9 expected transmembrane helices where the positive points were moved up by 0.02 to separate positives and negatives. Notice that the points for the negative set represent averages over the cross-validation models and therefore the predicted number of helices is not necessarily integer. The error bars show the standard deviation over the ten models.

Signal peptides and porins

The signal peptides that target a protein for export contain a hydrophobic region that can easily be mistaken for a transmembrane region by a prediction program. TMHMM was tested on a set of signal peptides by measuring how many of the signal peptides were predicted to be membrane proteins as described above. The result is shown in Table 3. For the eukaryotic and Gram-negative bacterial signal peptides, TMHMM erroneously identifies ~20 % as transmembrane helices. For signal peptides from Gram-positive bacteria, however, a full 60 % are predicted as transmembrane helices. Presumably, this is because signal peptides from Gram-positive bacteria have distinctly longer hydrophobic regions (von Heijne & Abrahmsen, 1989) than the other two classes.

Porins are the only class of membrane-spanning proteins besides helix-bundles known today. In these proteins, the membrane-spanning regions

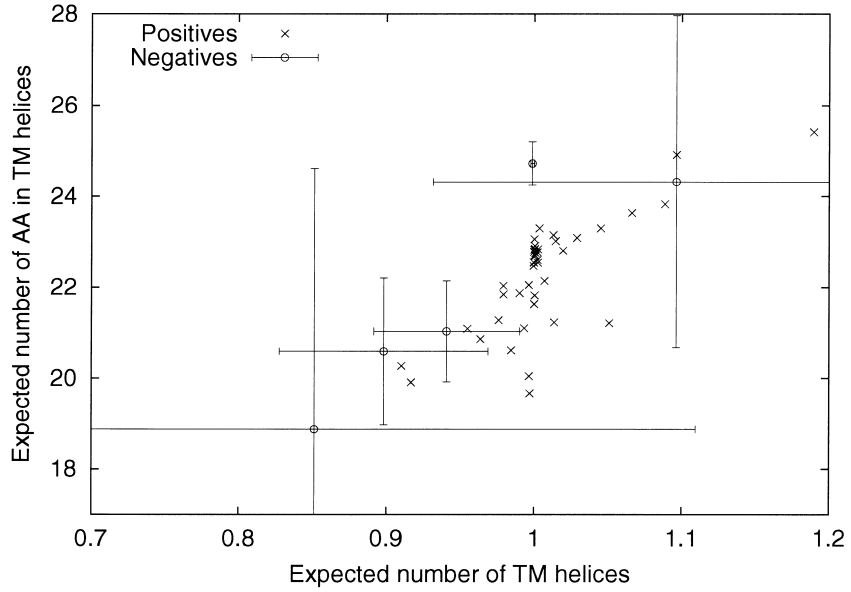

Figure 4. Correlation between the expected number of transmembrane helices and the expected number of amino acids in transmembrane helices. Only points around the critical region of about 0.9 expected transmembrane helices are shown. The error bars on the negative points show the standard deviation over the ten cross-validation models.

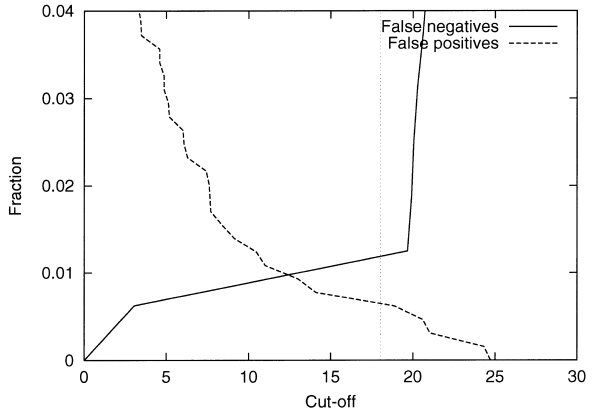

Figure 5. Discrimination between transmembrane proteins and soluble proteins. Discrimination based on the expected number of residues in transmembrane helices. The fraction of false negative (continuous line) and false positive predictions (broken line) as a function of the cut-off value. The broken line is at 18, which is the cut-off we used.

form a β-barrel (Cowan & Rosenbusch, 1994). Although their transmembrane regions generally are shorter and much less hydrophobic than those in helical membrane proteins, they could still be a source of falsely predicted transmembrane helices. After taking redundancy into account and removing structures that were genetically engineered variants of native porins, we were left with a porin set consisting of only six structures. These were analyzed in exactly the same way as the signal-peptide sets, and they were all predicted as not containing transmembrane helices. As expected, TMHMM thus does not identify porins as integral membrane proteins.

Genome-wide analysis of membrane proteins

Given that TMHMM is clearly superior to the prediction programs TOPPRED and ALOM (von Heijne, 1992; Klein *et al.*, 1985) both with regard to discrimination between membrane proteins and soluble proteins and with regard to topology prediction, we have repeated our earlier analysis (Wallin & von Heijne, 1998) of membrane proteins in organisms with fully sequenced genomes. We can now provide a better estimate of the number of membrane proteins in each organism, and also better estimates of the frequencies with which proteins of different topologies are found in different organisms.

A model was trained from all the 160 sequences in the training set according to the training scheme described below. After ascertaining that the number of false positives obtained with this model on the negative set was in agreement with the results above, this model was used for the genome studies. For each gene, the expected number of amino acid residues in transmembrane helices, as well as the topology prediction, were calculated. As shown above, signal peptides are sometimes falsely predicted as transmembrane helices. To correct this problem, we took all proteins with a predicted transmembrane helix at the N terminus that might be a signal peptide. These proteins were analyzed with SignalP-HMM (Nielsen & Krogh, 1998), and if a signal peptide was predicted, it was removed from the protein. This was done only for the eukaryotes and the Gram-positive and Gram-negative bacteria because SignalP is only developed for these groups of organisms (see Materials and Methods for details).

A preliminary test of the accuracy of SignalP-HMM reveals that about 80 % of the true signal peptides are found, and 20 % of transmembrane helices are mistaken for signal peptides in eukaryotes. For Gram-positive bacteria these estimates are 90 % and 10 %, and for Gram-negative, 95 % and 20 %. This test has not been properly cross-validated, but we believe that these numbers are reasonable estimates.

Estimates of the percentages of all annotated genes that encode integral membrane proteins of the helix bundle class are presented in Table 4. In general, these vary between ~20 % and ~30 %. A previously suggested correlation between the frequency of predicted membrane proteins and the total number of genes in the genome (Wallin & von Heijne, 1998) is not discernible in these new estimates (see Figure 6). Two organisms have noticeably higher fractions of membrane proteins

Table 2. False positives

PDB entry		Exp. number aa in membrane	Std. dev.
1RDZ (A)	Fructose 1,6-bisphosphatase;	24.3	3.6
1KVD (A)	Smk toxin	24.7	0.5
1NOX	Nadh oxidase	21.0	1.1
1CIY	CryIA(a) insecticidal toxin	20.6	1.6
1ENO	Enoyl acyl carrier protein reductase	18.9	5.7

The five proteins that are wrongly classified as transmembrane. They all have a known 3-D structure. The first column gives the PDB identifier with the chain in parenthesis. The second column gives the expected number of transmembrane helices averaged over the ten cross-validation models, and the last column the standard deviation. (Note that CryIA acts by inserting into the membrane, so it is perhaps not entirely a false positive.)

Table 3. The number of signal peptides predicted as transmembrane proteins

Class	No. of signal peptides	Predicted as tm protein
Eukaryotes	1011	209 (21 %)
Gram-negatives	266	60 (23 %)
Gram-positives	141	85 (60 %)

than the general trend: *Plasmodium falciparum* and *C. elegans*.

Representative plots of the number of proteins with a given predicted topology are shown in Figure 7. In general, the results are similar to that found in previous analyses (e.g. the high incidence of 12TM proteins in bacteria and of 7TM proteins in multi-cellular organisms), but we now see an additional feature not evident in earlier work: multi-spanning proteins with intracellular N and C termini are strongly preferred. From Table 5, it is clear that N_{in}-C_{in} topologies are 1.5-3 times more common than each of the other topologies. The only exception is *C. elegans*, where the 7TM proteins make the N_{out}-C_{in} topologies as common as the N_{in}-C_{in} topologies. To make sure that these results are not influenced by some systematic error in the TMHMM algorithm, we also used TOPPRED (von Heijne, 1992) to predict the topologies of the same proteins as used in the TMHMM analysis, and obtained essentially the same results (data not shown).

Discussion

TMHMM predicts transmembrane helices from single sequences with a high level of accuracy. Only about 2.5 % of the 696 helices in the data set of 160 proteins are missed, and about an equal number of false helices are predicted. About 77-78 % of the topologies are predicted correctly, and an additional 7 % were correct except that the topology was inverted, i.e. the cytoplasmic side was predicted as periplasmic and *vice versa*. This compares well to other methods, the best of which use multiple alignments to achieve the same level of accuracy (Rost *et al.*, 1996; Tusnady & Simon, 1998), see Sonnhammer *et al.* (1998) for comparisons. With our dataset and with no possibility to cross-validate, we measured a single-sequence accuracy of HMMTOP (Tusnady & Simon, 1998) of only 64 %. If we, however, remove known signal peptides from the sequences, the non-cross-validated accuracy of HMMTOP increases to 78 %. On the same data, TMHMM's cross-validated

Table 4. The number of predicted transmembrane proteins for several organisms

Organism	Number of annotated genes	Expected no AA > 18	One or more pred. TMHs	Reduced by signal peptides
S. cerevisiae	6305	1390 (22.05 %)	1303 (20.67 %)	50
C. elegans	19,099	5900 (30.89 %)	5778 (30.25 %)	285
D. melanogaster	14,100	2888 (20.48 %)	2835 (20.11 %)	106
A. thaliana (chrom. II and IV)	7859	1653 (21.03 %)	1578 (20.08 %)	217
P. falciparum (chrom. II and III)	225	98 (43.56 %)	91 (40.44 %)	2
E. coli	4289	910 (21.22 %)	898 (20.94 %)	135
H. influenzae	1709	328 (19.19 %)	323 (18.90 %)	48
H. pylori	1553	295 (19.00 %)	293 (18.87 %)	33
C. jejuni	1634	348 (21.30 %)	344 (21.05 %)	53
R. prowazekii	834	220 (26.38 %)	213 (25.54 %)	26
N. meningitidis	1989	352 (17.70 %)	354 (17.80 %)	38
M. tuberculosis	3918	747 (19.07 %)	691 (17.64 %)	95
B. subtilis	4100	983 (23.98 %)	987 (24.07 %)	145
M. genitalium	480	98 (20.42 %)	97 (20.21 %)	12
M. pneumoniae	677	126 (18.61 %)	122 (18.02 %)	23
T. pallidum	1031	241 (23.38 %)	244 (23.67 %)	-
B. burgdorferi	850	244 (28.71 %)	244 (28.71 %)	-
C. pneumoniae	1052	293 (27.85 %)	292 (27.76 %)	-
C. trachomatis	894	208 (23.27 %)	219 (24.50 %)	-
C. muridarum	818	189 (23.11 %)	198 (24.21 %)	-
A. aeolicus	1522	309 (20.30 %)	315 (20.70 %)	-
Synechocystis sp.	3169	816 (25.75 %)	818 (25.81 %)	-
D. radiodurans	3103	586 (18.88 %)	595 (19.17 %)	-
T. maritima	1846	422 (22.86 %)	445 (24.11 %)	-
M. jannashchii	1715	317 (18.48 %)	324 (18.89 %)	-
M. thermoautotrophicum	1869	407 (21.78 %)	407 (21.78 %)	-
A. fulgidus	2407	488 (20.27 %)	492 (20.44 %)	-
P. abyssi	1765	398 (22.55 %)	404 (22.89 %)	-
P. horikoshii	2064	567 (27.47 %)	534 (25.87 %)	-

For each organism the number of annotated genes is given, the number of predicted transmembrane proteins with the criterion that the most likely structure contains at least one transmembrane helix, and the number of predicted transmembrane proteins with the criterion that 18 or more residues are predicted to be in the membrane. Finally the number of predicted transmembrane proteins that were removed when correcting for signal peptides is given.

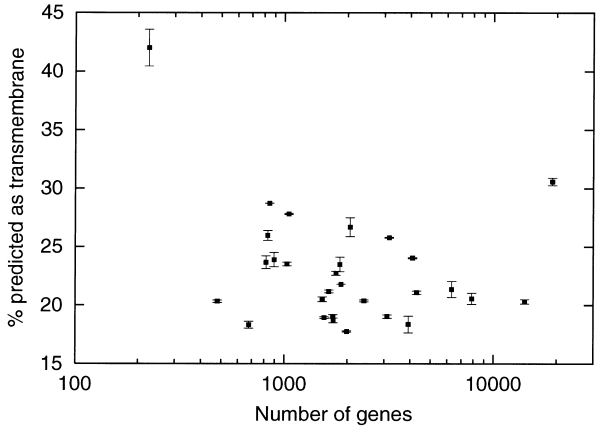

Figure 6. Number of predicted transmembrane proteins *versus* the number of genes. The errorbars show the two different predictions in Table 4.

accuracy is 79 %, while non-cross-validated TMHMM reaches 84 % accuracy.

In the analysis of mispredictions, it is interesting to note that the number of falsely merged helices and the number of split helices is very low (in fact zero). We believe this to be due to natural modeling of helix lengths and the grammar in the HMM. The main type of error made by TMHMM is to predict signal peptides as transmembrane helices for 60 % of Gram-positive bacterial proteins with a signal peptide, and for about 20 % of proteins from other organisms. Porins do not seem to be confused with helical membrane proteins.

There are several possible methods for discriminating between non-membrane and transmembrane proteins. Two expectation numbers that can be calculated from the model, the expected number of transmembrane helices and the expected number of amino acid residues in transmembrane helices, were shown to be equally good for discrimination. Using a set of 645 proteins with known 3D structure, it was shown that TMHMM discriminates very well between integral membrane proteins and water soluble proteins. Less than 1 % of the non-membrane proteins analyzed were wrongly classified, and only one of 160 membrane proteins was classified as non-membrane bound. However, the negative set used did not contain signal peptides.

We have compared the discriminative power of TMHMM to that of maxH (Boyd *et al.*, 1998), a recently developed method that has been optimized for discrimination but which cannot be used for topology prediction. The performance of TMHMM is slightly better in the sense that only six proteins are wrongly classified (one false negative plus five false positives) compared to ten for maxH (seven false negatives plus three false positives). The fact that maxH is more conservative in predicting transmembrane segments makes it better at discriminating signal peptides from transmembrane segments, however, by increasing the threshold for TMHMM the same effect could probably be achieved, but only at the price of a reduced accuracy in topology prediction. For this reason, we have chosen to deal with signal peptides by trying to remove them after the prediction of transmembrane helices.

With the TMHMM program, we estimate that integral membrane proteins of the helix bundle class account for roughly 20-30 % of all genes in most genomes, which was also found by Wallin & von Heijne (1998); only *P. falciparum* stands out as an exception, with around 40 % predicted membrane proteins, but this is based on a quite small number of genes annotated in chromosome 2 and 3. The two methods tested for discrimination (expected number of amino acid residues in the membrane and one or more predicted helices) yield almost identical results, within two percentage points except for *P. falciparum*, where the difference is about three points. Apart for the uncertainty in the predictions, there is also significant uncertainty in the annotated genes. The influence of wrongly annotated genes on these numbers is hard to estimate, but we believe it is small.

The problem with signal peptides is difficult to quantify exactly. Typically 10-25 % of the predicted transmembrane proteins were subjected to the analysis with SignalP. With an accuracy estimated to be around 20 %, it means that on the order of 2-5 % of the predictions have a ''signal peptide error''.

As noted in the Results, the possible correlation between the fraction of open reading frames (ORFs) encoding membrane proteins and genome size proposed earlier (Wallin & von Heijne, 1998) is not apparent in our new, more accurate estimates. The high fraction of membrane proteins in *C. elegans* is almost fully accounted for by the expansion of the G-protein coupled receptor family. Beyond this, however, the fraction of membrane proteins is remarkably constant between organisms.

An interesting new finding is that N_{in}-C_{in} topologies are preferred in all organisms except *C. elegans*. It has previously been noticed that N_{in} topologies are over-represented (Jones, 1998). All N_{in}-C_{in} proteins have an even number of transmembrane helices, and can be thought of as constructed from a succession of ''helical hairpins'', i.e. two transmembrane helices connected by an extracytoplasmic loop. Experimental studies have suggested that the helical hairpin may act as an independent ''insertion unit'' during membrane protein assembly (Gafvelin & von Heijne, 1994; Gafvelin *et al.*, 1997), and hence that topologies constructed from helical hairpin units may evolve more easily than other topologies. It is also clear from a number of experimental studies that the translocation of N-terminal tails across both the bacterial inner membrane and the ER membrane of eukaryotic cells places strong restrictions on the

Table 5. Statistics on the orientation of predicted membrane proteins

Organism	Number of annotated gens	Pred TMHs		Single spanning	Multispanning C_{in}	Multispanning C_{out}
S. cerevisiae	6305	1303*	N-term in	282	362	146
			N-term out	202	155	156
C. elegans	19,099	5778*	N-term in	1152	1074	495
			N-term out	919	1456	682
D. melanogaster	14,100	2835*	N-term in	692	650	263
			N-term out	502	371	357
A. thaliana (chrom. II and IV)	7859	1578*	N-term in	439	318	125
			N-term out	304	176	216
P. falciparum (chrom. II and III)	22	91*	N-term in	20	20	7
			N-term out	24	8	12
E. coli	4289	898*	N-term in	85	294	106
			N-term out	68	202	143
H. influenzae	1709	323*	N-term in	40	89	39
			N-term out	32	78	45
H. pylori	1553	293*	N-term in	48	78	23
			N-term out	40	53	51
C. jejuni	1634	344*	N-term in	54	89	39
			N-term out	35	76	51
R. prowazekii	834	213*	N-term in	49	49	29
			N-term out	18	39	29
N. meningitidis	1989	354*	N-term in	77	86	34
			N-term out	38	62	57
M. tuberculosis	3918	691*	N-term in	132	217	83
			N-term out	82	91	86
B. subtilis	4100	987*	N-term in	129	341	121
			N-term out	71	211	114
M. genitalium	480	97*	N-term in	9	25	9
			N-term out	18	22	14
M. pneumoniae	677	122*	N-term in	14	38	9
			N-term out	21	29	11
T. pallidum	1031	244	N-term in	83	73	28
			N-term out	15	19	26
B. burgdorferi	850	244	N-term in	96	60	24
			N-term out	13	26	25
C. pneumoniae	1052	292	N-term in	63	105	28
			N-term out	20	28	48
C. trachomatis	894	219	N-term in	49	78	27
			N-term out	16	20	29
C. muridarum	818	198	N-term in	49	71	21
			N-term out	14	19	24
A. aeolicus	1522	315	N-term in	71	91	39
			N-term out	33	41	40
Synechocystis sp.	3169	818	N-term in	242	213	98
			N-term out	62	87	116
D. radiodurans	3103	595	N-term in	118	185	77
			N-term out	40	93	82
T. maritima	1846	445	N-term in	135	133	55
			N-term out	33	41	48
M. jannashchii	1715	324	N-term in	77	103	35
			N-term out	29	53	27
M. thermoautotrophicum	1869	407	N-term in	92	109	60
			N-term out	46	63	37
A. fulgidus	2407	492	N-term in	93	168	69
			N-term out	53	60	49
P. abyssi	1765	404	N-term in	71	149	62
			N-term out	24	59	39
P. horikoshii	2064	534	N-term in	106	176	69
			N-term out	50	70	63

The number of genes and the number of predicted transmembrane proteins are shown first. The star indicates that the prediction has been corrected for signal peptides. Then follows the number of single-spanning and the number of multi-spanning predicted with each of the possible orientations.

amino acid sequence of the tail (Monne *et al.*, 1999; Whitley *et al.*, 1995, 1994), thus working against the appearance of N_{out} topologies during evolution.

TMHMM is available as a prediction server at http://www.cbs.dtu.dk/services/TMHMM. There are also pointers to the data sets used and other resources at this web site.

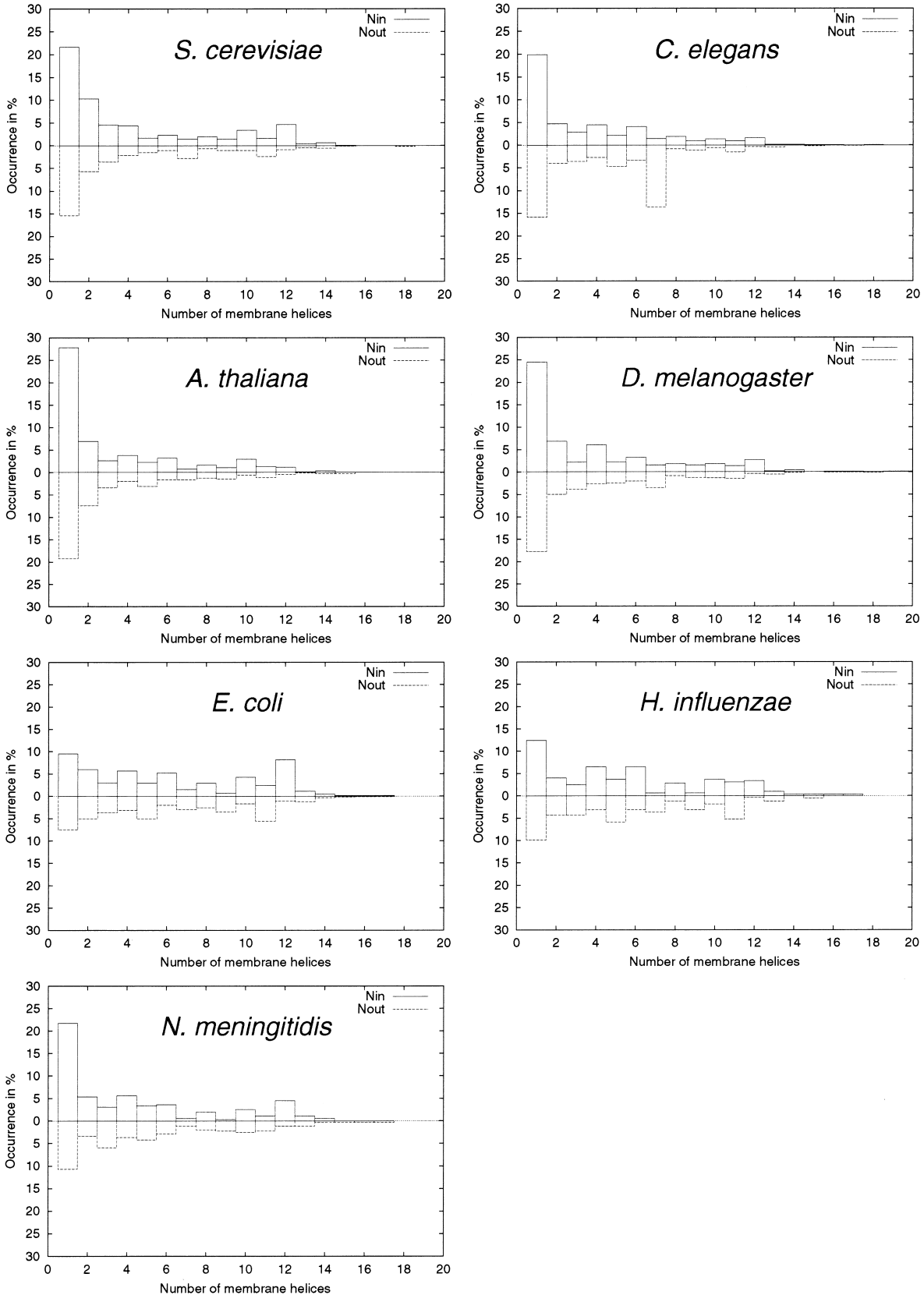

Figure 7 (*legend opposite*)

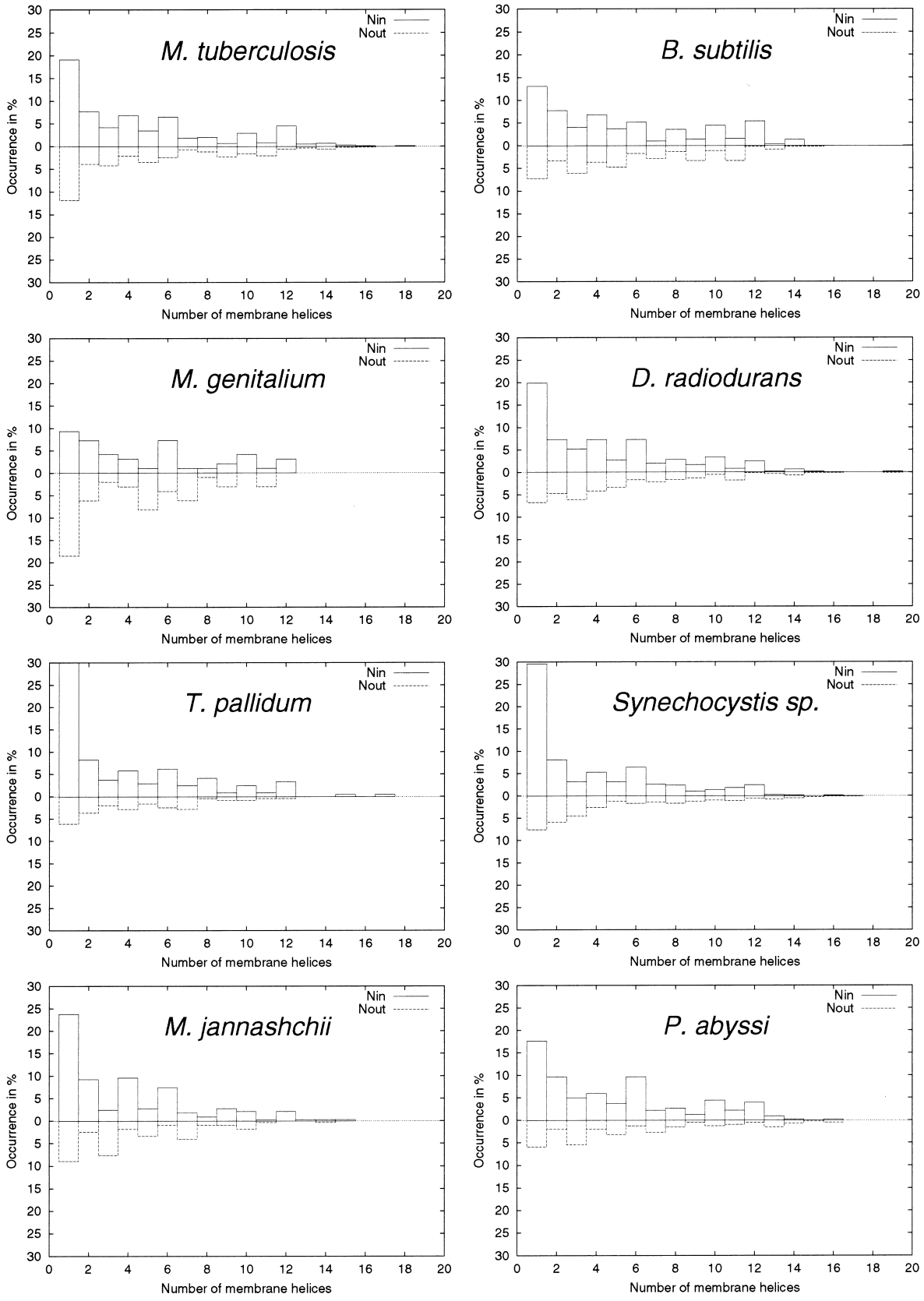

Figure 7. Distribution of transmembrane protein topologies. Histograms of the fraction of transmembrane proteins (predicted number/total number of annotated genes) with a certain topology for some of the genomes. The histograms for the ones with the N-terminal predicted outside are shown upside-down with a broken line.

Materials and Methods

Data sets

We have compiled a set of 160 proteins, most of which have experimentally determined topology. It contains 108 multi-spanning and 52 single-spanning proteins.

It should be noted that nearly all proteins with an experimentally determined topology have been analyzed with biochemical and genetic methods that are not always entirely reliable (Traxler *et al.*, 1993). Only a very small number of membrane protein structures have been determined at atomic resolution, and even in these cases the exact location of the membrane is not obvious (Wallin *et al.*, 1997). Given the uncertainty in the currently available data, perfect prediction accuracy is thus unrealistic. To avoid errors as much as possible, we did not include proteins for which different experiments have yielded conflicting topologies and where it was not obvious which topology was closer to the truth. Signal peptides were not removed from any of these proteins.

The dataset is available at http://www.cbs.dtu.dk/krogh/TMHMM/ in the ten cross-validation partitions (see below).

To test the discrimination between membrane proteins and other proteins, a set of 645 soluble proteins with known structure was used. This set has been extracted from the Protein Data Bank (PDB) and homologous sequences removed as described by Lund *et al.* (1997). The set is available at the above mentioned web address.

A set of signal peptides used for training of SignalP (Nielsen *et al.*, 1997) was used to test the discrimination between signal peptides and membrane helices. This redundancy-reduced set contains 1011 eukaryotic, 141 Gram-positive, and 266 Gram-negative sequences. The sequences consist of the signal peptides and 30 amino acid residues after the cleavage site. This set is available from ftp://virus.cbs.dtu.dk/pub/signalp/.

For porins, we used only crystallographically determined structures. Since most porin structures in PDB are annotated ''engineered'', very few native structures were found. The PDB entries 1mal, 1mpm, 1 pho, 1prn, 2omf and 2por were selected to serve as a porin test set.

All genes annotated in the Genbank entry of the genomes and chromosomes used were downloaded from ftp://ncbi.nlm.nih.gov/genbank/genomes/, except for *C. elegans*, which was downloaded from the URL: ftp://genome.wustl.edu/pub/gsc1/C_elegans/elegans.gz.

Cross-validation

Because of the lack of independent test data, all results reported are based on tenfold cross-validation. The set of 160 membrane proteins was partitioned into ten subsets with 16 proteins in each. It was made sure that no two proteins from different sets were more than 25% identical in a Needleman-Wunsch alignment by the ALIGN program in the FASTA package (Pearson, 2000). Within the sets, the similarity was allowed to be higher. Cross-validation was done by training on all sequences in nine subsets, and testing the accuracy on the subset left out from training. This was repeated for all ten subsets.

For the discrimination between membrane proteins (positives) and water soluble proteins (negatives) the fraction of false positives of the entire negative set was found for all ten models at a given cut-off value, and the average reported. The fraction of false negatives was found by adding up the false negatives for each of the models tested on the corresponding test set and dividing by the size of the entire set (160).

HMM training

Usually HMMs are estimated by maximizing the likelihood $P(x^1, \ldots, x^N|\theta)$ of the sequences $x^1, \ldots, x^N$ in the training set with respect to the model parameters θ. For the transmembrane model used here, it would require that the different segments (helices, inside loops, etc) be cut out of the sequences and the corresponding submodels estimated separately. Instead, we use labeled sequences for the estimation (Durbin *et al.*, 1998; Krogh, 1994, 1997), which is a simple generalization of the standard method that constrains residues labeled as membrane helix to use only states labeled the same, and those labeled as cytoplasmic to used states for cytoplasmic residues, etc. The HMM is estimated in three stages.

In the first stage, the aim is to correct for the inaccurate boundaries of the annotated transmembrane helices. This is done by allowing six residues around the boundaries to match any state in the model. For instance, at a cytoplasmic boundary of a helix we start with a labeling ''MMMMMMiiiiii'' for the 12 residues around the boundary (M represents membrane helix and i represents inside or cytoplasmic). Now we put wildcard labels (.) around the boundary, so the labels become 'MMM.....iii'. During estimation the wildcards can match any state, so in this case they can match states in both the membrane helix submodel and in the submodel for inside residues. If a loop is shorter than six, it is ensured that the middle label (or the two middle ones) always remains, so for instance 'MMMMMiiiiMMMMM' becomes 'MM....ii....MM'. The choice of three residues to each side seems to work well, but has not been seriously optimized.

During model estimation, the boundaries were then placed automatically within the allowed window so as to optimize the total likelihood of the model. We use the Baum-Welch iterative re-estimation procedure, which is guaranteed to converge to a local maximum of the likelihood. It is well known that there is a problem with many suboptimal local maxima of the likelihood for an HMM. Therefore, noise was added to the model parameters during the estimation procedure, but the amount of noise was decreased in each iteration of the procedure until it reached zero, after which point the estimation procedure is continued until convergence. To be precise: to a model parameter p, the amount of noise added is $A * p * n$, where n is a random number between 0 and 1, and A is the amplitude of the noise, which starts at $A = 1$ and is multiplied by 0.8 in each iteration of the estimation procedure. It has been shown that such a procedure improves on the final likelihood (Hughey & Krogh, 1996), although it may not lead to the global maximum.

In the second stage of estimation, the helix boundaries were re-estimated with the first model. This was done by again ''unlabeling'' the helix boundaries (this time by five residues to each side), and then finding the most probable labeling constrained by the remaining labels. It means that the over-all topology of the protein was fixed, but the model decided where to put the helix boundaries within a window of ten residues. After the relabeling, a new model was estimated where all labels

are fixed. This estimation started from the model of stage one and was estimated with the Baum-Welch procedure without noise.

In the third and final stage of estimation, the model from stage two was further optimized using a discriminative method of estimation as described by Sonnhammer *et al*. (1998).

Prediction

To find the most probable topology of a membrane protein, we used the N-best algorithm described by Krogh (1997).

The posterior probability that a residue is found in a helix, the cytoplasm, or the periplasm is calculated in the following way. The posterior probability of each model state is found by the forward-backward algorithm (Durbin *et al*., 1998; Rabiner, 1989). These probabilities are then added up for states belonging to each of the three categories. These probabilities are used for plots like Figure 2.

The expected number of residues in a transmembrane helix is found by simply adding the posterior probabilities for each of the residues being in a transmembrane helix. To find the expected number of helices in a protein, we picked a state that had to be used for an outgoing helix, and the corresponding state for an ingoing helix. The posterior probabilities (as found above) for these two states were then added along the sequence. This gives the expected number of times the sequence passes through a helix.

The analysis of each genome was done in the following way. The expected number of residues in the membrane was calculated for each protein. If it was larger than three the number of predicted transmembrane helices was found by the N-best algorithm (this prefiltering was done to save computer time). If the organism was a eukaryote or a Gram-positive or negative bacterium, it was checked for signal peptides. Proteins with a transmembrane helix predicted less than 50 amino acid residues from the N terminus, and an N terminus predicted as inside, were extracted as likely candidates for signal peptides. Such proteins were sent to SignalP-HMM (http://www.cbs.dtu.dk/services/SignalP-2.0/), and if a cleavage site was predicted with a probability of more than 0.5, the predicted signal peptide was cleaved off. Then the prediction of transmembrane helices was redone, with the change that the prediction was constrained to have the N terminus outside. These predictions were used for all statistics.

For comparison the maxH program 'New_-maxH_v3.3.pl' was downloaded from http://beck2.-med.harward.edu/resources/maxh/ and ran on a Unix workstation with default settings.

Acknowledgments

This work was supported by a grant from the Danish National Research Foundation to A.K., and by grants from the Swedish Technical Sciences Research Council and the Foundation from Strategic Research to G.v.H.

References

Argos, P., Rao, J. K. & Hargrave, P. A. (1982). Structural prediction of membrane-bound proteins. *Eur. J. Biochem.* **128**, 565-575.

Bowie, J. U. (1997). Helix packing in membrane proteins. *J. Mol. Biol.* **272**, 780-789.

Boyd, D., Schierle, C. & Beckwith, J. (1998). How many membrane proteins are there? *Protein Sci.* **7**, 201-205.

Casadio, R., Fariselli, P., Taroni, C. & Compiani, M. (1996). A predictor of transmembrane alpha-helix domains of proteins based on neural networks. *Eur. Biophys. J.* **24**, 165-178.

Cowan, S. W. & Rosenbusch, J. P. (1994). Folding pattern diversity of integral membrane proteins. *Science,* **264**, 914-916.

Durbin, R. M., Eddy, S. R., Krogh, A. & Mitchison, G. (1998). *Biological Sequence Analysis*, Cambridge University Press, Cambridge, UK.

Edelman, J. (1993). Quadratic minimization of predictors for protein secondary structure. *J. Mol. Biol.* **232**, 165-191.

Gafvelin, G. & von Heijne, G. (1994). Topological ''frustration'' in multispanning *E. coli* inner membrane proteins. *Cell,* **77**, 401-412.

Gafvelin, G., Sakaguchi, M., Andersson, H. & von Heijne, G. (1997). Topological rules for membrane protein assembly in eukaryotic cells. *J. Biol. Chem.* **272**, 6119-6127.

von Heijne, G. (1986). The distribution of positively charged residues in bacterial inner membrane proteins correlates with the trans-membrane topology. *EMBO J.* **5**, 3021-3027.

von Heijne, G. (1992). Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. *J. Mol. Biol.* **225**, 487-494.

von Heijne, G. (1994). Membrane proteins: from sequence to structure. *Annu. Rev. Biophys. Biomol. Struct.* **23**, 167-192.

von Heijne, G. (1999). A day in the life of Dr K, or how I learned to stop worrying and love lysozyme: a tragedy in six acts. *J. Mol. Biol.* **293**, 367-379.

von Heijne, G. & Abrahmsen, L. (1989). Species-specific variation in signal peptide design. Implications for protein secretion in foreign hosts. *FEBS Letters,* **244**, 439-446.

Hughey, R. & Krogh, A. (1996). Hidden Markov models for sequence analysis: extension and analysis of the basic method. *CABIOS,* **12**, 95-107.

Jones, D. T. (1998). Do transmembrane protein superfolds exist? *FEBS Letters,* **423**, 281-285.

Jones, D. T., Taylor, W. R. & Thornton, J. M. (1994). A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry,* **33**, 3038-3049.

Klein, P., Kanehisa, M. & DeLisi, C. (1985). The detection and classification of membrane-spanning proteins. *Biochim. Biophys. Acta,* **815**, 468-476.

Klemm, P., Tong, S., Nielsen, H. & Conway, T. (1996). The gntP gene of *Escherichia coli* involved in gluconate uptake. *J. Bacteriol.* **178**, 61-67.

Krogh, A. (1994). Hidden Markov models for labeled sequences. In *Proceedings of the 12th IAPR International Conference on Pattern Recognition*, pp. 140-144, Los AlamitosIEEE Computer Society Press, California.

Krogh, A. (1997). Two methods for improving performance of a HMM and their application for gene

finding. In *Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology* (Gaasterland, T., Karp, P., Karplus, K., Ouzounis, C., Sander, C. & Valencia, A., eds), pp. 179-186, AAAI Press, Menlo Park, CA.

Lund, O., Frimand, K., Gorodkin, J., Bohr, H., Bohr, J., Hansen, J. & Brunak, S. (1997). Protein distance constraints predicted by neural networks and probability density functions. *Protein Eng.* **10**, 1241-1248.

Monne, M., Gafvelin, G., Nilsson, R. & von Heijne, G. (1999). N-tail translocation in a eukaryotic polytopic membrane protein: synergy between neighboring transmembrane segments. *Eur. J. Biochem.* **263**, 264-269.

Nielsen, H. & Krogh, A. (1998). Prediction of signal peptides and signal anchors by a hidden Markov model. In *Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology* (Glasgow, J., Littlejohn, T., Major, F., Lathrop, R., Sankoff, D. & Sensen, C., eds), pp. 122-130, AAAI Press, Menlo Park, CA.

Nielsen, H., Engelbrecht, J., Brunak, S. & von Heijne, G. (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* **10**, 1-6.

Pearson, W. R. (2000). Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol. Biol.* **132**, 185-219.

Persson, B. & Argos, P. (1994). Prediction of transmembrane segments in proteins utilising multiple sequence alignments. *J. Mol. Biol.* **237**, 182-192.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE,* **77**, 257-286.

Rost, B., Casadio, R., Fariselli, P. & Sander, C. (1995). Transmembrane helices predicted at 95% accuracy. *Protein Sci.* **4**, 521-533.

Rost, B., Fariselli, P. & Casadio, R. (1996). Topology prediction for helical transmembrane proteins at 86% accuracy. *Protein Sci.* **5**, 1704-1718.

Sonnhammer, E. L., von Heijne, G. & Krogh, A. (1998). A hidden Markov model for predicting transmembrane helices in protein sequences. In *Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology* (Glasgow, J., Littlejohn, T., Major, F., Lathrop, R., Sankoff, D. & Sensen, C., eds), pp. 175-182, AAAI Press, Menlo Park, CA.

Traxler, B., Boyd, D. & Beckwith, J. (1993). The topological analysis of integral cytoplasmic membrane proteins. *J. Membr. Biol.* **132**, 1-11.

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.

Wallin, E. & von Heijne, G. (1998). Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci.* **7**, 1029-1038.

Wallin, E., Tsukihara, T., Yoshikawa, S., von Heijne, G. & Elofsson, A. (1997). Architecture of helix bundle membrane proteins: an analysis of cytochrome c oxidase from bovine mitochondria. *Protein Sci.* **6**, 808-815.

Whitley, P., Zander, T., Ehrmann, M., Haardt, M., Bremer, E. & von Heijne, G. (1994). Sec-independent translocation of a 100-residue periplasmic N-terminal tail in the *E. coli* inner membrane protein proW. *EMBO J.* **13**, 4653-4661.

Whitley, P., Gafvelin, G. & von Heijne, G. (1995). SecA-independent translocation of the periplasmic N-terminal tail of an *Escherichia coli* inner membrane protein. Position-specific effects on translocation of positively charged residues and construction of a protein with a C-terminal translocation signal. *J. Biol. Chem.* **270**, 29831-29835.

Edited by F. Cohen

(Received 19 June 2000; received in revised form 15 October 2000; accepted 1 November 2000)