doi:10.1016/S0022-2836(03)00182-7 *J. Mol. Biol.* (2003) **327**, 735–744

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Reliability Measures for Membrane Protein Topology Prediction Algorithms

Karin Melén[1], Anders Krogh[2] and Gunnar von Heijne[1]*

[1]*Department of Biochemistry and Biophysics, Stockholm Bioinformatics Center Stockholm University SE-106 91 Stockholm, Sweden*

[2]*Department of Molecular Biology, Bioinformatics Centre University of Copenhagen Universitetsparken 15 DK-2100 Copenhagen Denmark*

**Corresponding author*

We have developed reliability scores for five widely used membrane protein topology prediction methods, and have applied them both on a test set of 92 bacterial plasma membrane proteins with experimentally determined topologies and on all predicted helix bundle membrane proteins in three fully sequenced genomes: *Escherichia coli*, *Saccharomyces cerevisiae* and *Caenorhabditis elegans*. We show that the reliability scores work well for the TMHMM and MEMSAT methods, and that they allow the probability that the predicted topology is correct to be estimated for any protein. We further show that the available test set is biased towards high-scoring proteins when compared to the genome-wide data sets, and provide estimates for the expected prediction accuracy of TMHMM across the three genomes. Finally, we show that the performance of TMHMM is considerably better when limited experimental information (such as the in/out location of a protein's C terminus) is available, and estimate that at least ten percentage points in overall accuracy in whole-genome predictions can be gained in this way.

Keywords: membrane protein; topology prediction; bioinformatics

Introduction

It is estimated that some 20–25% of all open reading frames (ORFs) in fully sequenced genomes encode integral membrane proteins.[1] Strikingly, however, considerably less than 1% of all 3D protein structures deposited in the Protein Data Bank[2] are of membrane proteins. Theoretical structure prediction methods are thus of particular importance for membrane proteins. Most current methods in this field do not deal with predicting the 3D structure, but rather try to predict the most likely topology of the protein, i.e. the in/out location of the N and C termini relative to the membrane, and the number and positions of the membrane-spanning regions. Topology information can be generated experimentally by different approaches such as gene fusion, proteolytic digestion *in situ*, antibody binding, and chemical modification. A good topology model is a necessary prerequisite for experimental structure–function studies and can be used as a starting point for attempts to model the 3D structure.

From a structural point of view, there are two major groups of integral membrane proteins: the helix bundle proteins, in which one or several α-helices span the membrane, and the β-barrel proteins, in which eight or more anti-parallel transmembrane β-strands form a closed barrel. The β-barrel membrane proteins have so far been found only in the outer membranes of Gram-negative bacteria, mitochondria, and chloroplasts, whereas the α-helical membrane proteins are present in all types of membranes. Here, we consider only methods for predicting the topology of helix bundle membrane proteins.

The best current topology prediction methods are claimed to predict the correct topology for some 70–85% of all proteins, although, as will be shown below, this is an overestimate. Rather, we estimate an overall prediction accuracy of 55–60% correctly predicted topologies when entire proteomes are analyzed. Importantly, none of the most widely used methods (except PHD, see below) provides any estimate of the reliability of a given prediction, i.e. some measure of whether the topology of a particular protein is more or less likely to be correct than average.

In this study, we have tried to construct useful reliability scores for five widely used topology prediction methods: TMHMM,[1] HMMTOP,[3] MEMSAT,[4] PHD[5] and TopPred.[6] The goal has been to use these scores to compare performance characteristics on a test set of proteins with

Abbreviations used: ORF, open reading frame.

E-mail address of the corresponding author: gunnar@dbb.su.se

experimentally determined topologies with performance characteristics on three complete genomes, *Escherichia coli*, *Saccharomyces cerevisiae* and *Caenorhabditis elegans*, and to assess to what extent limited, easily obtainable experimental topology information can be used to improve the theoretical predictions.

Results

Construction of reliability scores

Judging by published bench-marking studies, TMHMM, HMMTOP and MEMSAT seem to have the best overall performance characteristics of the available topology prediction programs.[7,8] Two less well performing but widely used methods, PHD and TopPred, have been included for comparison. Each method is described below in some detail, together with a discussion of the reliability scores that we have constructed from the raw output from each program.

TMHMM

TMHMM is based on a hidden Markov model with seven types of states (helix core, helix caps on either side of the membrane, short loop on cytoplasmic side/inside, short and long loop on non-cytoplasmic side/outside, and a globular domain state). Each type of state has a probability distribution over the 20 amino acids that have been estimated from membrane proteins with experimentally known topologies. TMHMM outputs the most probable topology of the protein given the model. The output is a labelled sequence of the three classes i (inside or cytoplasmic), h (helix) and o (outside or extra-cytoplasmic) that obeys the "biological grammar" that a helix must be followed by a loop and that inside and outside loops must alternate. Posterior probabilities for being in the three classes (p(i), p(h), and p(o)) are calculated for every residue in the sequence. We have constructed three different reliability scores ($S1$–$S3$) for TMHMM (see Methods).

S1: The mean posterior probability of the labelled sequence. A high mean posterior probability indicates that most of the residues have a high probability for their assigned classes and thus that the overall prediction might be considered reliable. The posterior probability values for each residue are calculated as described.[1] A possible shortcoming of this score is that a small region with low probabilities embedded in a long sequence with generally high scores will not greatly affect $S1$, even though it indicates an uncertainty in the prediction.

S2: The minimum posterior probability in the sequence of labelled residues. A low $S2$ score indicates that there is at least one part of the protein where the prediction is doubtful. Since the probability values close to the borders between different classes often are low, even though the exact point of transition between one class and another generally makes no difference to the overall topology, we mask out a small number of residues (three, five, seven, nine) on each side of each border before locating the minimum probability value. For the score evaluation presented below, we masked out nine residues at each side of each border; the results are essentially the same in the whole interval three to nine masked residues (data not shown).

S3: The quotient p(best topology)/p(all possible topologies), calculated after a masking step as described below. The two probability values are included in the standard TMHMM output, where p(best topology) is calculated with the N-best algorithm and p(all possible topologies) is calculated with the forward algorithm, as described.[1] A quotient close to 1 implies that the best path through the model (i.e. the predicted topology) is much more probable than all alternative paths (i.e. all other topologies). TMHMM can generate a list of several high-scoring paths where the top ones frequently have very similar topologies (corresponding to shifts of one or a few residues at the borders between different classes that do not change the overall topology). Since the exact borders between the classes are not generally known even for the experimentally determined topologies, it is reasonable to mask out some residues (we have used ten) on each side of a class border and consider all topologies compatible with the "best" topology after masking as the same prediction. We thus sum the probabilities for all paths that give the same topology prediction after masking as the best path before dividing by p(all possible paths) as obtained from the raw output.

HMMTOP

HMMTOP is a hidden Markov model with five states (inside loop, inside helix tail, helix, outside helix tail and outside loop). For a given amino acid sequence it finds the most probable path through the model. Instead of taking into account only the absolute amino acid composition in the separate parts of the protein, it searches for the combination of states that gives the highest difference in the amino acid distributions. The idea is that a switch in the topology should be reflected in a large amino acid distribution change (maximum divergence). In the raw output, numbers are given for the entropy of the best path (i.e. the most probable topology) and the entropy of the whole model. We have used the difference in entropy (i.e. entropy of best path − entropy of model) as a measure of the reliability. The smaller the difference, the better the best path represents the whole model, and the more likely to be correct the predicted topology should be.

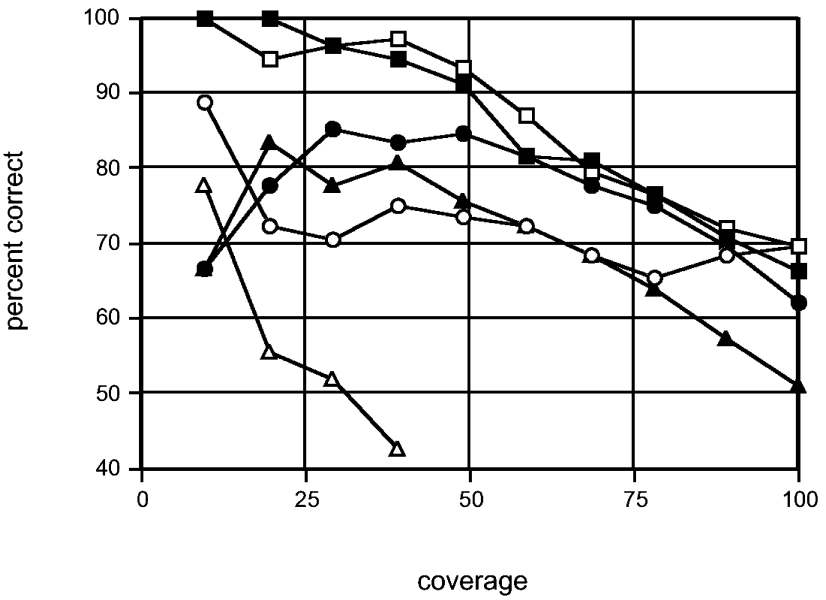

Figure 1. Relation between test set cumulative coverage and the fraction of correct topology predictions for five different prediction methods over a set of 92 prokaryotic membrane proteins with experimentally determined topologies. TMHMM *S*3 score, filled squares; MEMSAT, open squares; HMMTOP, open circles; PHDhtm (web version, multi-sequence mode), filled circles; PHDhtm (single-sequence mode), filled triangles; TopPred, open triangles (for TopPred, many sequences did not generate more than one topology. For those cases no reliability score could be calculated, which explains the total TopPred coverage of only 36%).

MEMSAT

MEMSAT is based on a model with five structural states (inside loop, inside helix end, helix middle, outside helix end, outside loop). Each state is associated with a statistical table (log likelihoods) of the frequency of the 20 amino acids. The tables have been constructed from membrane proteins of known topologies and treat single- and multispanning membrane proteins separately. A dynamic programming algorithm solves the problem of finding the optimal state assignments for the query sequence. The algorithm computes scores for all possible topologies starting with one helix, and then increases the number of helices one at a time until the scores become too low. The output produces a list of topologies representing all possible number of TM helices (in both orientations) and their scores. The topology with the highest score is the final prediction. To assess the reliability, we have calculated the difference in scores between the best and the second best prediction. If the difference is high, the top-scoring topology should be more likely to be correct.

PHD

PHD is a general tool for predicting secondary structure of proteins, and the PHDhtm routine is the part handling membrane proteins. It is designed to use information from homologous proteins. The first step in the method is a BLAST search[9] against the SWISSPROT database.[10] A multiple sequence alignment of the hits is constructed and a neural network then estimates the preference for each residue to be in a transmembrane helix or in a loop. The highest-scoring putative transmembrane segment is used in a second step to decide whether the protein is a helix bundle integral membrane protein. The third step is a dynamic program algorithm that finds the optimal number and locations of transmembrane regions (the model). Finally, the overall orientation of the protein in the membrane is predicted by applying the "positive-inside" rule.[11,12]

PHDhtm is the only method in our study that automatically provides some sort of reliability measure. In the output, there is one reliability index for the model (i.e. for the number and locations of the transmembrane regions) that is based on a comparison between the two highest-scoring models, and a second reliability index for the orientation that is proportional to the charge difference between the outside and inside parts of the protein. Both indices range from 0 (low) to 9 (high). However, the two indices are not combined into a single reliability score for the overall topology. We have evaluated both the two existing indices and the mean value of the two indices as reliability scores.

Because the other four methods only use information in a single query sequence (and not information from homologous sequences) we decided to run PHDhtm in single-sequence mode for the main analysis. However, we have also used the multi-sequence mode for comparison.

TopPred

TopPred was the first topology prediction method that combined hydrophobicity analysis and the positive-inside rule. It first calculates a standard hydrophobicity profile for the query protein. Peaks above an upper cut-off (i.e. regions rich in hydrophobic residues) are considered to be confident transmembrane helix predictions whereas peaks between the upper and a lower cut-off are regarded as putative transmembrane helices. Consequently, several topologies can be constructed with or without the putative helix/helices. Out of these possible topologies, the one with the largest difference in the number of positively charged amino acids between the two

sides of the membrane is given as the best prediction. We have calculated a reliability score as the difference between the charge-difference values for the two top-scoring topologies. If no putative helices are identified from the hydrophobicity plot, only one topology is predicted, and thus no reliability score can be calculated in such cases.

Reliability scores correlate with prediction accuracy

The five methods and their corresponding reliability scores described above were evaluated over a previously collected test set (see Methods) composed of 92 prokaryotic helix bundle membrane proteins with experimentally determined topologies. For each method and score, the 92 topology predictions were ranked from high to low scores. The results are summarized in Figure 1 in the form of a plot of prediction accuracy *versus* cumulative coverage of the test set.

As is clear from this Figure, TMHMM and MEMSAT have the best prediction characteristics according to this test (for TMHMM, only the *S*3 score is shown, as the *S*1 and *S*2 scores yield essentially the same results). For both methods, ~50% of the predictions have reliability scores corresponding to a prediction accuracy of ~90%, and ~70% of the proteins have scores corresponding to a prediction accuracy ~80%. If the entire test set is considered (100% coverage), the prediction accuracy is 65–70%.

For HMMTOP, PHDhtm, and TopPred, our definitions of reliability scores do not seem very useful. We repeated the PHDhtm analysis by running the web version in multi-sequence mode, which improved the overall accuracy on the whole test set from 51% to 63%, but did not improve the discrimination between good and bad predictions based on the reliability score. The two individual reliability indices given by PHDhtm were no better than the mean reliability score shown in the Figure (data not shown).

Interestingly, the top-scoring proteins are, to a significant extent, different for the two best methods, TMHMM and MEMSAT. By simply combining the two scores as shown in Figure 2 (TMHMM score $S3 > 0.7$ and/or MEMSAT score >4) we reach a prediction accuracy of ~95% for the ~60% top-scoring proteins in the test set. However, this apparent improvement needs to be confirmed on a larger data set. A more elaborate scheme for combining different topology prediction methods has been presented,[8] and it is possible that one can find "optimized" combinations of reliability scores that perform better than the individual scores discussed here.

Proteins with known topology constitute a biased set compared to full-size proteomes

The development and evaluation of topology prediction methods is, to some extent, limited by the available experimental data, and a significant fraction of the proteins in our test set have been used in the original construction of the different prediction methods. This has made it difficult to obtain realistic estimates of the expected performance characteristics when the methods are applied to previously uncharacterized proteins, and different authors come to different conclusions on this point.[7,8] From a couple of recent studies,[13,14] it is clear, however, that the available test sets of proteins with experimentally determined topologies is biased, although the extent of the bias is unknown.

The reliability scores constructed here make it possible to address this question using whole-genome data. We have therefore calculated the TMHMM *S*3 score distributions for the predicted helix bundle membrane protein proteomes of one prokaryotic, *E. coli*[15] and two eukaryotic, *S. cerevisiae*[16] and *C. elegans*,[17] organisms, and have compared these distributions to the distributions obtained for the test set.

As TMHMM has been shown to be able to discriminate between soluble and integral membrane proteins with very great accuracy,[1] the three membrane protein proteomes were defined as all ORFs for which TMHMM predicts at least two transmembrane helices. Predicted single-spanning proteins were not included, since cleavable signal peptides are often predicted as transmembrane helices, thus erroneously identifying many secreted proteins as single-spanning membrane proteins. Even so, an unknown proportion of the membrane proteins identified in this way will contain cleavable signal peptides, in contrast to the test set proteins, which all lack cleavable signal

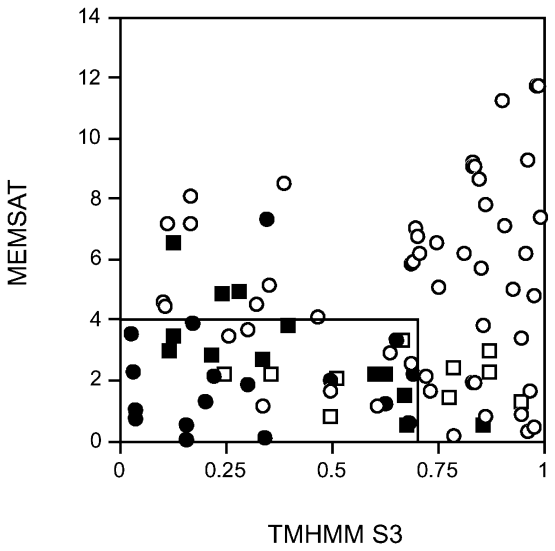

Figure 2. TMHMM *S*3 and MEMSAT scores for 92 test set proteins. Open circles, both predictions correct; filled circles, both predictions false; open squares, TMHMM prediction correct, MEMSAT prediction false; filled squares, TMHMM prediction false, MEMSAT prediction correct.

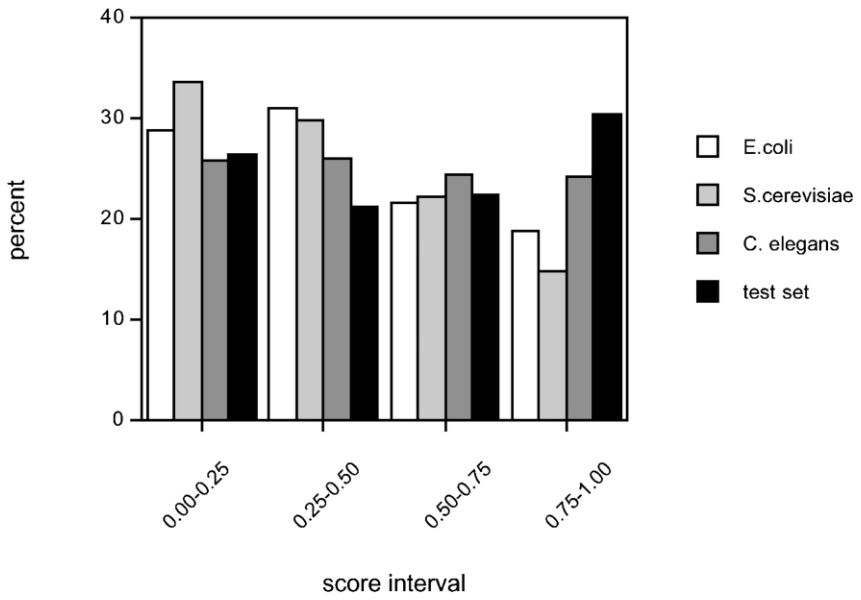

Figure 3. TMHMM *S*3 score distributions. The fraction of all predicted membrane proteins with two or more TM helices in each genome or in the test set (76 proteins) and for each score interval is shown.

peptides. This may reduce the *S*3 scores slightly for some of the predicted proteins, but we consider it unlikely that this is enough to explain the differences between the proteome sets and the test set reported below.

The results are presented in Figure 3, where the percentages of membrane proteins are plotted for different score intervals. To be able to compare the score distributions for the three proteomes with the test set, we removed all single-spanning sequences in the test set, ending up with 76 sequences and a TMHMM accuracy for this reduced set of 63%. The most striking result is that there is a much larger fraction of high-scoring proteins in the test set compared to the three proteomes, and thus that the overall prediction accuracy of ~66% reported in Figure 1 is a clear overestimate. To obtain a more realistic estimate, we first derived an empirical relation between the prediction accuracy and the *S*3 score by dividing the 92 test set predictions, ranked from high to low scores, into four equal-size groups and then plotting the average prediction accuracy in each group against the mean score for that group, Figure 4(A). The accuracy/score relation is reasonably well described by the straight line $A = 80 \times S3 + 20$. Using this relation, we calculated the expected *A*-values for all proteins in the respective membrane protein proteomes, which is plotted against the cumulative coverage in Figure 4(B). As a control, we plotted the real mean accuracy and the calculated accuracy (*A*) for the test set; the two latter curves agree well and we thus conclude that the expected accuracy *A* is a reasonable representation of the real data. The mean prediction accuracies estimated in this way for the whole proteomes (56% for *E. coli*, 53% for *S. cerevisiae* and 59% for *C. elegans*) are significantly lower than the ~66% obtained for the test set, suggesting that the widely quoted prediction accuracies of 70–85% are serious overestimates.

There are several possible explanations for the test set bias. First, even though jack-knife procedures were used in the development of the prediction methods, there are many subtle ways in which the methods may have been overtrained. It is quite likely that the proteins for which experimental topologies have been reported have some characteristics such as unusually hydrophobic transmembrane segments that simultaneously simplify both experimental mapping and prediction. There are many families of membrane proteins for which no experimental topology is available and which have thus not been seen by the prediction methods.

Looking more carefully at the results for the individual genomes (Figure 3), it is interesting to note that *S. cerevisiae* has a particularly large fraction of low-scoring proteins, while *C. elegans* and *E. coli* have more similar score distributions. We did not expect *C. elegans* to have the greatest predicted accuracy, since it is a eukaryote and the relationship $A = 80 \times S3 + 20$ was derived from prokaryotic proteins. However, we suspected that the family of 7TM-receptors, known to be exceptionally large in *C. elegans*,[18] might have contributed to the results. We therefore identified all *C. elegans* proteins predicted to have seven transmembrane helices and an extracellular N terminus (985 out of totally 4059) and analyzed the 7TM and non-7TM sets separately. The 7TM set was found to have a score distribution similar to that of the test set, whereas the score distribution for the remaining *C. elegans* membrane proteins almost coincided with that of *E. coli* (data not shown). Finally, the combination of the TMHMM *S*3 and MEMSAT scores discussed above (Figure 2), gave the following coverages for the three proteomes: 45% for *E. coli*, 46% for *S. cerevisiae* and 56% for *C. elegans*,

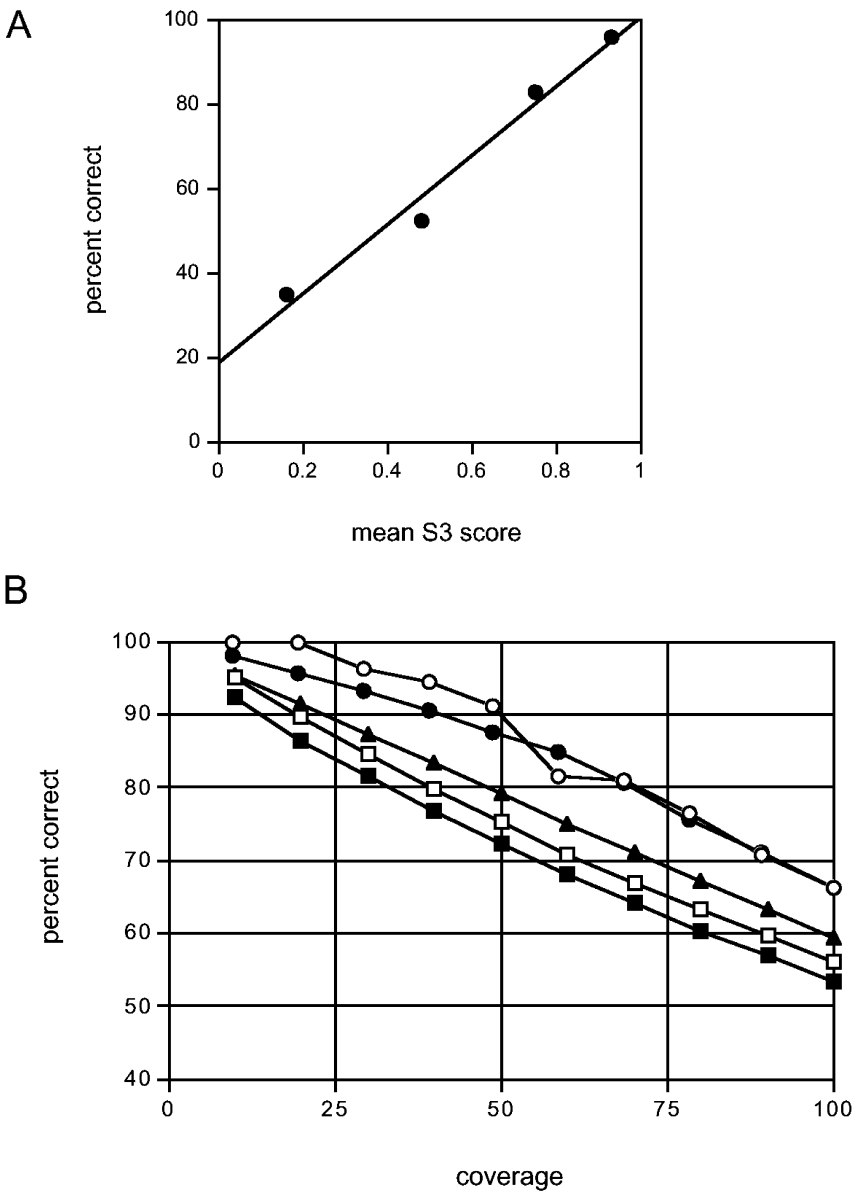

Figure 4. Expected performance of TMHMM over all predicted membrane proteins with two or more TM helices in each genome. (A) Mean fraction of correctly predicted proteins *versus* the mean TMHMM *S*3 score for each quartile of the test set of 92 proteins. The least-squares fit is given by $A = 80 \times S3 + 20$, where A is the expected accuracy (i.e. the probability that a prediction with a given *S*3 score is correct). (B) Estimated relation between cumulative coverage and the fraction of correct topology predictions for the test set of 92 proteins and for all predicted membrane proteins with two or more TM helices in each genome. Test set (original data), open circles; test set (calculated data), filled circles; *C. elegans*, filled triangles; *E. coli*, open squares; *S. cerevisiae*, filled squares.

which should be compared to the 60% coverage of the test set.

Inclusion of limited experimental information: a strategy for large-scale topology mapping

Given the rather low estimates for the expected mean prediction accuracy over full-size proteomes discussed above, it is clear that topology predictions, in general, provide only a rough guide to the true topology of a protein. On the other hand, the reliability scores presented here can be used to reduce considerably the necessary experimental work required to reach a satisfactory level of prediction accuracy.

We have shown that limited experimental information such as a determination of the in/out location of the C terminus of a protein can be used in conjunction with topology prediction to rapidly provide a very reliable topology model, at least in certain cases.[19] With the introduction of reliability scores, it is now possible to extend this strategy to entire proteomes. The basic TMHMM algorithm allows one to fix the class-assignment for any position in the sequence by setting the probability for a position to belong to a certain class to 1.0 *a priori*. If the C-terminal residue of each protein in the test set is assigned to its experimentally known class, the relation between accuracy and coverage becomes much more favourable and the overall mean accuracy increases from 66% to 77% (Figure 5(B)). Similarly, if the N terminus is fixed, the overall mean accuracy increases to 79%, and if both termini are fixed it reaches 88% (data not shown). Again, there is an approximately linear relationship between the accuracy and the *S*3 score; with a fixed C terminus, the relation is $A^c = 70 \times S3^c + 30$ (data not shown).

Finally, we tried to estimate how much the prediction accuracy across the *E. coli*, *S. cerevisiae* and

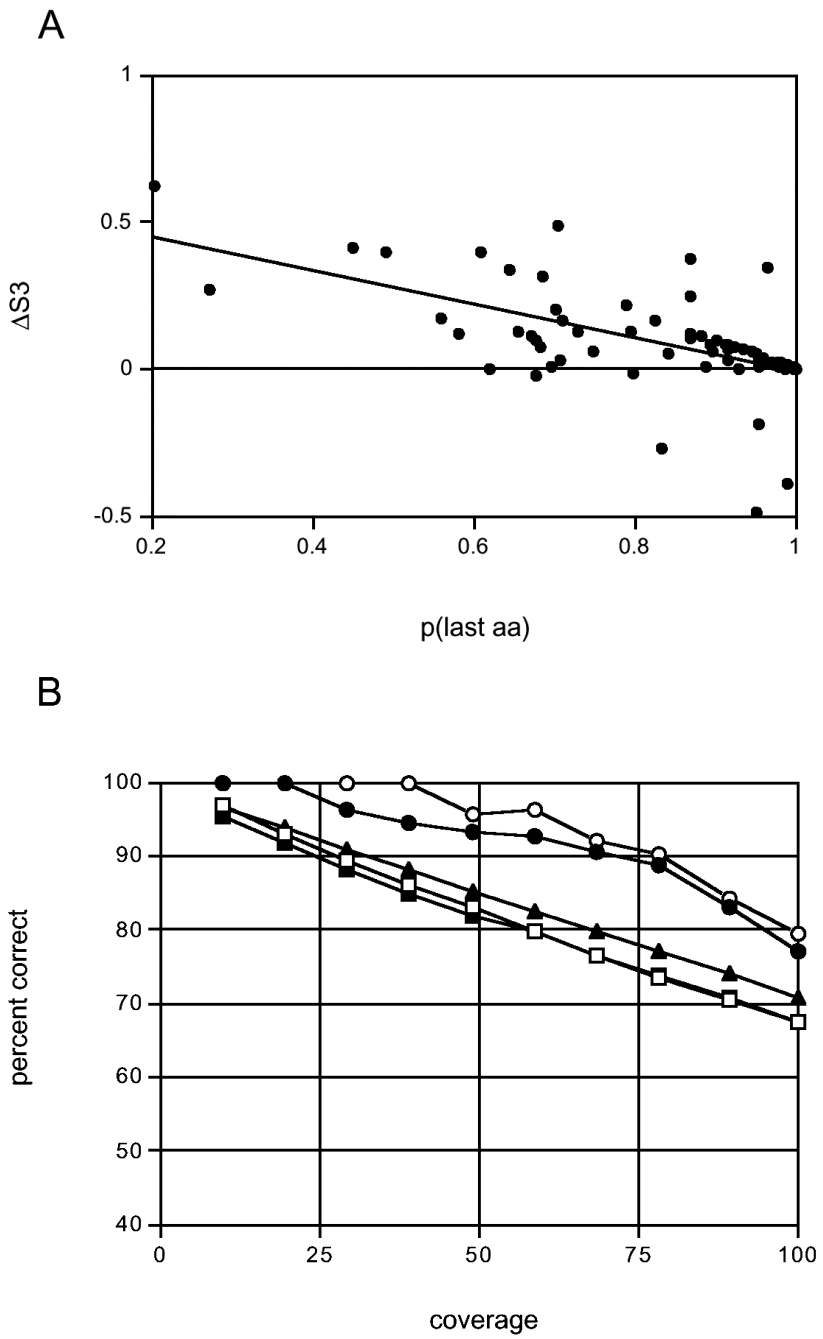

Figure 5. Influence of experimental information on TMHMM performance. (A) Relation between increase in *S*3 score for the test set of 92 proteins with the C-terminal residue fixed to its experimentally known location and the value of *p*(last aa); $\Delta S3^{c} = -0.57 \times p(\text{last aa}) + 0.57$. (B) Relation between cumulative coverage and fraction of correct predictions. Observed accuracy for the test set with fixed C-terminal locations, filled circles; and with fixed N-terminal locations, open circles. Expected accuracy, A^{c*}, for the three genomes assuming that the C-terminal location is known: *C. elegans*, filled triangles; *E. coli*, open squares; *S. cerevisiae*, filled squares.

C. elegans membrane protein proteomes would improve if the location of each protein's C terminus was known. To this end, we used the test set to measure the difference in *S*3 score, $\Delta S3^{c}$, between the score obtained with the C terminus fixed and the score obtained in the absence of any experimental information ($\Delta S3^{c} = S3^{c} - S3$) and plotted $\Delta S3^{c}$ *versus* the probability value for the location of the C-terminal residue obtained in the absence of experimental information, *p*(last aa), i.e. the probability value for the assigned class of the last amino acid in the sequence (Figure 5(A)). Although the data are rather scattered, there is a linear trend described by $\Delta S3^{c} = -0.57 \times p(\text{last aa}) + 0.57$. In other words, the smaller the value of *p*(last aa), the larger is the mean increase in the *S*3 score when the C-terminal residue is assigned to its known class. This expression was used for estimating the increase in *S*3 score for all proteins in the three proteomes from which the estimated $S3^{c}$ scores can be calculated; $S3^{c*} = S3 + \Delta S3^{c}$, assuming that the C-terminal location is known. The expected accuracy, A^{c*}, was then calculated from the expression for A^{c} above. The results are shown in Figure 5(B). The estimated increase in overall accuracy for the proteomes is from 56% to 67% for *E. coli*, from 53% to 67% for *S. cerevisiae*, and from 59% to 71% for *C. elegans*. It should be emphasised that these numbers are only rough estimates, but they nevertheless suggest that

prediction performance would improve significantly if C-terminal mapping data were available.

Generally applicable methods for determining the location of the C-terminal end of a protein on the basis of either reporter fusions or engineered acceptor sites for N-linked glycosylation exist for *E. coli*, *S. cerevisiae* and mammalian proteins,[20–22] and we have shown that such methods can be used on a relatively large scale (our unpublished work). On the basis of TMHMM-predictions,[1] we have estimated that the membrane protein proteome of *E. coli* consists of 769 proteins with two or more transmembrane helices, and that of *S. cerevisiae* of 847 such proteins. The results presented above suggest that highly reliable topology models for a majority of these proteins should be obtainable from a simple experimental determination of the C-terminal location.

Discussion

Membrane protein topology prediction is an important area in contemporary bioinformatics, and provides a useful starting point for experimental studies of membrane proteins. While the overall performance of different topology prediction methods has been much discussed lately,[7,8,13] essentially no work has been done trying to estimate the reliability of individual predictions. Here, we have constructed simple reliability scores for five widely used methods, TMHMM, HMMTOP, MEMSAT, PHDhtm and TopPred, and have applied them to a test set of 92 prokaryotic proteins with experimentally determined topologies and to the full-size membrane protein proteomes from *E. coli*, *S. cerevisiae* and *C. elegans*.

For TMHMM and MEMSAT, there is a good correlation between the reliability scores we have defined and the expected accuracy of a prediction. For both methods, ~50% of the predictions have reliability scores corresponding to a prediction accuracy of ~90%, and ~70% of the proteins have scores corresponding to a prediction accuracy of ~80% over the test set. For the remaining three methods, we were unable to derive useful reliability scores.

We have further used the TMHMM reliability score to assess the degree of bias in the test set as compared to the predicted membrane protein proteomes of *E. coli*, *S. cerevisiae* and *C. elegans*. In conformity with the results of two recent studies,[13,14] we find that the test set is biased towards high-scoring proteins, and we estimate that only some 53–59% of all predicted topologies for these proteomes are correct, compared to 63% for the test set when only proteins with two or more transmembrane helices are considered (or 66% for the whole test set). The reliability scores make it possible to estimate the likelihood that a given prediction is correct, allowing experimental topology mapping efforts to be focused on proteins with low reliability scores.

Finally, we have tried to estimate the expected improvement in prediction accuracy if the in/out location of the C terminus of every protein in a proteome was known from experimental data, since relatively rapid methods for such determinations are now available. For all three proteomes, we find that TMHMM will predict the correct topology for ~70% of all membrane proteins, given that the C-terminal location is known. Again, the likelihood that a given prediction is correct can be estimated from the reliability score.

In summary, we describe new reliability scores for TMHMM and MEMSAT, two of the currently best-performing topology prediction methods, that make it possible to estimate the likelihood that a given prediction is correct and that can be used in conjunction with limited experimental information to provide high-quality topology models for entire proteomes.

Methods

Prediction methods

TMHMM2.0,[1] HMMTOP2.0,[3,23] MEMSAT version 1.8,[4] PHDhtm version 1998.01[5] and TopPred version 1.0[6] were used in single-sequence mode and with default parameter settings. PHDhtm was also run in its multiple sequence alignment mode on the website.†

Definition of reliability scores

$$\text{TMHMM } S1 = (p_1(\text{label}) + p_2(\text{label}) + \cdots + p_N(\text{label}))/N$$

where N is the sequence length and $p_i(\text{label})$ is the posterior probability for the assigned class (label = i, o or h) for residue i.

$$\text{TMHMM } S2 = \min[p_1(\text{label}), p_2(\text{label}), \ldots, p_N(\text{label})]$$

$$\text{TMHMM } S3 = p(\text{best topology})/p(\text{all possible topologies})$$

To calculate p(best topology) we first identify all high-scoring predictions that are compatible with the highest-scoring one by masking ten residues on either side of each class border. All predictions that have the same class assignments as the highest-scoring one after masking are considered as being the same, and p(best topology) is the summed probabilities (as given by TMHMM) for these predictions. These individual probabilities as well as p(all possible topologies) are calculated as described.[1]

$$\text{HMMTOP} : \text{score} = \text{entropy(best path)} - \text{entropy(model)}$$

$$\text{MEMSAT} : \text{score} = \text{score(best topology)} - \text{score(second best topology)}$$

$$\text{PHDhtm} : \text{score} = ((\text{index(model)} + \text{index(orientation)}))/2$$

† http://cubic.bioc.columbia.edu/predictprotein/

TopPred : score

$$= \Delta \text{ positive charges(best topology)}$$

$$- \Delta \text{ positive charges(second best topology)}$$

Definition of correct predictions

A predicted topology is considered correct if it has the correct number of transmembrane segments and the correct location of the N terminus.

Data sets

The test set used is a collection of 92 prokaryotic helix bundle membrane proteins with experimentally known topologies.[24] We selected proteins belonging to "trust levels" A, B and C, but excluded level C proteins with only partial topologies. We removed all sequences that were annotated to contain an N-terminal signal or a pro-peptide.

The highest level of sequence identity (as determined by ClustalW[25] alignments) between any two proteins in the test set was 59%, and 71 sequences had less than 30% mutual identity as determined by the Hobohm 2 algorithm.[26]

For the proteome analysis, all predicted ORFs from three fully sequenced genomes, *E. coli*†, *S. cerevisiae*‡ and *C. elegans*§, were downloaded.

To extract the membrane proteins, TMHMM was run on all ORFs in the respective genomes and all proteins with two or more predicted transmembrane segments were retained. Proteins with a single predicted transmembrane segment were not included, since a considerable but unknown fraction of these segments are cleavable signal peptides rather than transmembrane helices.[1] The numbers of proteins analyzed were 749 for *E. coli*, 847 for *S. cerevisiae* and 4059 for *C. elegans*.

Acknowledgements

This work was supported by a grant from the Swedish Knowledge Foundation *via* the Research School of Medical Bioinformatics and AstraZeneca to K.M., and by grants from the Foundation for Strategic Research and the Swedish Research Council to G.v.H.

References

1. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. (2001). Predicting transmembrane protein topology with a hidden Markov model. Application to complete genomes. *J. Mol. Biol.* **305**, 567–580.
2. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H. *et al.* (2000). The Protein Data Bank. *Nucl. Acids Res.* **28**, 235–242.
3. 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.
4. 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.
5. Rost, B., Fariselli, P. & Casadio, R. (1996). Topology prediction for helical transmembrane proteins at 86% accuracy. *Protein Sci.* **5**, 1704–1718.
6. von Heijne, G. (1992). Membrane protein structure prediction—hydrophobicity analysis and the positive-inside rule. *J. Mol. Biol.* **225**, 487–494.
7. Möller, S., Croning, M. & Apweiler, R. (2001). Evaluations of methods for the predictive evaluation of membrane spanning regions. *Bioinformatics*, **17**, 646–653.
8. Ikeda, M., Arai, M., Lao, D. & Shimizu, T. (2001). Transmembrane topology prediction methods: a re-assessment and improvement by a consensus method using a data-set of experimentally characterized transmembrane topologies. *In Silico Biol.* **2**, 1–15.
9. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.
10. O'Donovan, C., Martin, M. J., Gattiker, A., Gasteiger, E., Bairoch, A. & Apweiler, R. (2002). High-quality protein knowledge resource: SWISS-PROT and TrEMBL. *Brief. Bioinformat.* **3**, 275–284.
11. 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.
12. von Heijne, G. (1989). Control of topology and mode of assembly of a polytopic membrane protein by positively charged residues. *Nature*, **341**, 456–458.
13. Käll, L. & Sonnhammer, E. (2002). Reliability of transmembrane predictions in whole-genome data. *FEBS Letters*, **532**, 415–418.
14. Nilsson, J., Persson, B. & von Heijne, G. (2002). Prediction of partial membrane protein topologies using a consensus approach. *Protein Sci.*, **11**, 2974–2980.
15. Blattner, F. R., Plunkett, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M. *et al.* (1997). The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1462.
16. Goffeau, A., Aert, R., Agostini-Carbone, M., Ahmed, A., Aigle, M., Alberghina, L. *et al.* (1997). The yeast genome directory. *Nature*, suppl. **387**, 1–105.
17. Stein, L., Sternberg, P., Durbin, R., Thierry-Mieg, J. & Spieth, J. (2001). WormBase: network access to the genome and biology of *Caenorhabditis elegans*. *Nucl. Acids Res.*, **29**, 82–86.
18. Bargmann, C. (1998). Neurobiology of the *Caenorhabditis elegans* genome. *Science*, **282**, 2028–2033.
19. Drew, D., Sjöstrand, D., Nilsson, J., Urbig, T., Chin, C. N., de Gier, J. W. & von Heijne, G. (2002). Rapid topology mapping of *Escherichia coli* inner-membrane proteins by prediction and PhoA/GFP fusion analysis. *Proc. Natl Acad. Sci. USA*, **99**, 2690–2695.
20. Manoil, C. (1991). Analysis of membrane protein topology using alkaline phosphatase and β-galactosidase gene fusions. *Methods Cell. Biol.* **34**, 61–75.

† http://bmb.med.miami.edu/EcoGene/EcoWeb/
‡ ftp://genome-ftp.stanford.edu/pub/yeast/yeast_ORFs/
§ ftp://ftp.sanger.ac.uk/pub/wormbase/

21. Deak, R. & Wolf, D. (2001). Membrane topology and function of Der3/Hrd1p as a ubiquitin-protein ligase (E3) involved in endoplasmic reticulum degradation. *J. Biol. Chem.* **276**, 10663–10669.
22. Popov, M., Tam, L. Y., Li, J. & Reithmeier, R. A. F. (1997). Mapping the ends of transmembrane segments in a polytopic membrane protein—scanning N-glycosylation mutagenesis of extracytosolic loops in the anion exchanger, Band 3. *J. Biol. Chem.* **272**, 18325–18332.
23. Tusnady, G. E. & Simon, I. (2001). The HMMTOP transmembrane topology prediction server. *Bioinformatics*, **17**, 849–850.
24. Möller, S., Kriventseva, E. & Apweiler, R. (2000). A collection of well-characterised integral membrane proteins. *Bioinformatics*, **16**, 1159–1160.
25. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* **22**, 4673–4680.
26. Hobohm, U., Scharf, M., Schneider, R. & Sander, C. (1992). Selection of representative protein data sets. *Protein Sci.* **1**, 409–417.

Edited by F. E. Cohen

(Received 19 November 2002; received in revised form 30 January 2003; accepted 31 January 2003)