BIOINFORMATICS ORIGINAL PAPER

Vol. 23 no. 5 2007, pages 538–544
doi:10.1093/bioinformatics/btl677

*Sequence analysis*

# Improving the accuracy of transmembrane protein topology prediction using evolutionary information

David T. Jones

Department of Computer Science, University College London, Gower Street, London WC1E 6BT, United Kingdom

Received on October 4, 2006; revised on January 3, 2007; accepted on January 4, 2007

Advance Access publication January 19, 2007

Associate Editor: Thomas Lengauer

**ABSTRACT**

**Motivation:** Many important biological processes such as cell signaling, transport of membrane-impermeable molecules, cell–cell communication, cell recognition and cell adhesion are mediated by membrane proteins. Unfortunately, as these proteins are not water soluble, it is extremely hard to experimentally determine their structure. Therefore, improved methods for predicting the structure of these proteins are vital in biological research. In order to improve transmembrane topology prediction, we evaluate the combined use of both integrated signal peptide prediction and evolutionary information in a single algorithm.

**Results:** A new method (MEMSAT3) for predicting transmembrane protein topology from sequence profiles is described and benchmarked with full cross-validation on a standard data set of 184 transmembrane proteins. The method is found to predict both the correct topology and the locations of transmembrane segments for 80% of the test set. This compares with accuracies of 62–72% for other popular methods on the same benchmark. By using a second neural network specifically to discriminate transmembrane from globular proteins, a very low overall false positive rate (0.5%) can also be achieved in detecting transmembrane proteins.

**Availability:** An implementation of the described method is available both as a web server (http://www.psipred.net) and as downloadable source code from http://bioinf.cs.ucl.ac.uk/memsat. Both the server and source code files are free to non-commercial users. Benchmark and training data are also available from http://bioinf.cs.ucl.ac.uk/memsat.

**Contact:** dtj@cs.ucl.ac.uk

## 1 INTRODUCTION

Integral membrane proteins mediate a wide range of fundamental biological processes such as cell signaling, transport of membrane-impermeable molecules, cell–cell communication, cell recognition and cell adhesion. Not surprisingly, therefore, understanding the structure and function of membrane proteins is of great importance in biological and pharmacological research.

Analysis of the complete genomic sequences for several organisms indicates that 20–25% of all genes code for transmembrane proteins (Jones, 1998; Wallin and von Heijne, 1998). Despite this, only around 1% of all 3-D protein structures deposited in the Protein Data Bank are of membrane proteins (Berman *et al.*, 2000), mainly because they are difficult to crystallize and are generally too large for study by NMR. This provides a strong impetus for the development of efficient structure prediction methods for transmembrane proteins.

The main principle underlying the structure and stability of membrane proteins is the high energetic cost of dehydrating the peptide bond during its transfer into a non-polar environment (Bowie, 2005; White, 2001). This has two consequences. Firstly, and perhaps most obviously, most of the amino acid side chains found within transmembrane segments should be non-polar. Secondly, the polar groups of the main chain of transmembrane segments must 'self-satisfy' their own hydrogen bonding potential in order to lower the energetic cost of membrane insertion. This second constraint is typically accomplished by exploiting two structural motifs: the first is the membrane-spanning alpha-helix and the second is the beta-barrel (White and Wimley, 1999).

Although there has been some recent progress in predicting the full 3-D structure of transmembrane proteins (e.g. Yarov-Yarovoy *et al.*, 2006), the most widely applied prediction technique for these proteins is to determine the transmembrane topology, i.e. the inside–outside location of the N and C termini relative to the cytoplasm, along with the the number and sequence locations of the membrane spanning regions. Knowing a membrane protein's topology can be a significant step toward inferring both its structure and function.

Transmembrane protein topology prediction methods rely on two major topological features. The first is that, as already discussed, transmembrane helices are generally formed by hydrophobic sequence stretches; the second is the bias towards positively charged residues in the regions flanking the hydrophobic stretches, especially on the intracellular side of the membrane. The feature is commonly known as 'the positive-inside rule', where short loops are found to be enriched with Lys and Arg residues on the intracellular side and depleted on the outside (von Heijne, 1992; Wallin and von Heijne, 1998).

Many methods have been developed for predicting the topology of helix-bundle membrane proteins, and here we review a few of the better known and representative methods. For a more extensive recent review of progress, see Chen *et al*., (2002) and Fleishman and Ben-Tal (2006) for a detailed analysis of methods published prior to 2002.

In the earliest transmembrane prediction methods, simple hydrophobicity plots were used (e.g. Kyte and Doolitle, 1982) to detect probable transmembrane segments. Kyte and Doolitle used a 'sliding window' approach to identify membrane segments where a fixed window of width 19 residues was moved along the protein sequence and the average hydrophobicity was calculated for amino acids within the window. Using these mean hydrophobicity values, a threshold could be identified for deciding whether the centre of the window is within a membrane–spanning membrane helix or not. The Kyte and Doolitle method, along with other similar approaches, only identifies likely transmembrane segments, and the approach is not able to predict the inside–outside phasing of the segments relative to the cytoplasm.

The first attempt at a method for predicting transmembrane topology was the TopPred method proposed by von Heijne (1992). TopPred again makes use of hydrophobicity plots to predict transmembrane segments, but combines these predictions with the 'positive-inside rule' mentioned earlier. The observation that there was a strong bias for positively charged residues on the inside-facing segments of a transmembrane protein provided a means for identifying which predicted topology was correct from a small number of alternatives.

The MEMSAT method of Jones *et al*. (1994) was the first prediction method to fully integrate the prediction of transmembrane topology with the prediction of transmembrane segments. Rather than simply deciding between a few possible topological models, MEMSAT was able to calculate the most probable length, location and topological orientation for each transmembrane segment, guaranteeing a mathematically optimal solution. The method made use of scores compiled from membrane protein data and a dynamic programming algorithm to search through all possible topological models by a process of expectation maximization. The propensity of each amino acid to be in one of five states (inside loop, outside loop, inside helix end, helix middle and outside helix end) was calculated from experimentally well-described membrane proteins and was expressed as a log-likelihood ratio. This approach can be seen as a forerunner of more recent approaches based on hidden markov models. MEMSAT2 (Jones, 1998) made use of the same scoring tables as MEMSAT and the same dynamic programming algorithm, but used sequence profiles to produce a consensus topology score across an aligned family of sequences.

PHDhtm (Rost *et al*., 1996) was the first method to use neural networks for the prediction of transmembrane protein structure. It used multiple sequence alignments to do a consensus prediction of transmembrane segments in the target protein, and then predicted the overall topology by applying the positive-inside rule.

TMHMM (Sonnhammer *et al*., 1998) and HMMTOP (Tusnady and Simon, 1998) were the first methods to employ hidden markov models to the problem of transmembrane topology prediction. TMHMM implements a cyclic model with seven states for transmembrane helix, whereas HMMTOP uses hidden Markov models to distinguish between five structural states [helix core, inside loop, outside loop, helix caps (C and N) and globular domains]. The states are connected by transition probabilities. As with the earlier MEMSAT approach, dynamic programming is used to match a sequence against the model in order to find the most probable topology.

Finally, in line with developments in methods for predicting globular protein structure, a few consensus methods have been proposed. The first such approach has been developed by Nilsson *et al*. (2002), which uses the consensus of five topology prediction methods (TMHMM, HMMTOP, MEMSAT, PHD, TopPred). They find that for the 50% of *Escherichia coli* proteins where high agreement is observed between methods, the topology is found to agree with experimental evidence around 90% of the time. Another approach is to make a variety of predictors available via a single Web site to allow users to use their own judgment in deciding a consensus prediction (Amico *et al*., 2006).

In this paper, the widely used MEMSAT method has been combined with a neural network trained on sequence profiles. This should allow the method to directly take into account sequence conservation information that has proven to be very powerful in globular protein secondary structure prediction, as demonstrated by the success of the PSIPRED method (Jones, 1999). To date, few methods have made use of multiple sequence alignments in making transmembrane topology predictions. Martelli *et al*. (2003) made use of a neural network along with two HMM predictors. The neural network in this case only has a single output, which indicates whether the target residue is in a TM segment or not. Topology prediction is handled by a set of *ad hoc* rules. In contrast to this, here we use a neural network to predict not only TM segments but also to score the topology and to identify possible signal peptides.

Another interesting recent attempt at using multiple-sequence information is PolyPhobius (Käll *et al*., 2005), based on their original Phobius method (Käll *et al*., 2004), which makes use of an HMM to predict both transmembrane topology and signal peptide cleavage sites. In PolyPhobius, a multiple-sequence alignment is used to calculate the best 'average' path through the states of the HMM. Some improvement in prediction accuracy is observed over the single-sequence approach, but clearly a more explicit treatment of evolutionary information is needed to provide significant improvements in accuracy.

## 2 METHODS

The first step in the MEMSAT3 algorithm makes use of a feed-forward neural network comprising 399 inputs, 15 hidden units and 4 output units. The inputs correspond to a window size of 19 residue positions and 21 inputs per residue. This encoding is the same as that used by the PSIPRED secondary structure prediction method (Jones, 1999), though with a slightly longer window in this case. A number of different output encodings were considered, but it was decided that a minimal encoding of just four outputs would be preferable for optimal neural network training. The four output targets are: cytoplasmic ($O_{\mathrm{in}}$), non-cytoplasmic ($O_{\mathrm{out}}$), transmembrane segment ($O_{\mathrm{tm}}$) and signal peptide ($O_{\mathrm{sig}}$).

As a testing and training set, the widely used data set of Möller *et al*. (2001) was used. This comprises 188 proteins where experimental evidence exists for the given topology and includes 10 artificially mutated *E.coli* leader peptidase sequences that provide a particular challenge to topology prediction methods. Of the 188 proteins in the original set, four were excluded. Entry 60IM_ECOLI had no TM segments annotated. Entry MEL_APIME is a bilayer disrupting peptide found in bee venom and not a native integral membrane protein. Both CLC1_HUMAN and CITN_KLEPN were excluded because their topologies could not be reconciled with more recent annotations in SWISS-PROT (Boeckmann *et al.*, 2003). The final makeup of the final data set is shown in Table 1.

As with PSIPRED, PSI-BLAST (Altschul *et al.*, 1997) was used to calculate position-specific scoring matrices for each of the 184 proteins. As transmembrane proteins are prone to picking up false positives in PSI-BLAST searches, in this case only two iterations were used (–j 2) with a profile-inclusion E-value threshold of 0.001 (–h 0.001). In order to avoid partial sequences being included in the alignments, which can affect topology predictions, the SWISS-PROT sequence data bank was used, with any sequences labeled with keywords such as 'FRAGMENT' being excluded.

To allow proper cross-validation, separate training runs were carried out for every target protein, giving 184 sets of neural network weights in total. In training, the target sequence, along with any proteins in the training set found to be significantly similar, were excluded. The PSI-BLAST results were used for determining which sequences should be excluded, again using an E-value threshold of 0.001.

To calculate the most probable topology based on the neural network output, the MEMSAT dynamic programming algorithm (Jones *et al.*, 1994) was used. In the original MEMSAT method, five different regions of a transmembrane protein were defined as follows: inside-loop, inside-helix-cap, middle helix segment, outside-helix-cap and outside-loop. At every position in the target sequence, the four neural network outputs were combined as follows to generate scores for the MEMSAT topology search:

$$\begin{aligned} S_{\text{inside-loop}} &= 1/2\,(O_{\text{in}} - O_{\text{out}} - O_{\text{tm}}) \\ S_{\text{outside-loop}} &= 1/2\,(O_{\text{out}} - O_{\text{in}} - O_{\text{tm}}) \\ S_{\text{outside-tm}} &= 2(O_{\text{tm}} + O_{\text{out}} - O_{\text{in}}) \\ S_{\text{inside-tm}} &= 2(O_{\text{tm}} + O_{\text{in}} - O_{\text{out}}) \\ S_{\text{middle-tm}} &= 2\,O_{\text{tm}} \end{aligned}$$

Where $O_x$ represents the raw neural network output $x$. For evaluating segments of the target sequence within the first 25 residues of the target protein, the signal-peptide output is subtracted from the inside-loop term and added to the outside-loop term as follows:

$$\begin{aligned} S_{\text{inside-loop}} &= 1/2\,(O_{\text{in}} - O_{\text{out}} - O_{\text{tm}}) - O_{\text{sig}} \\ S_{\text{outside-loop}} &= 1/2\,(O_{\text{out}} - O_{\text{in}} - O_{\text{tm}}) + O_{\text{sig}} \end{aligned}$$

In this way, predicted signal peptides biased predictions towards having a non-cytoplasmic N-terminus.

In order to avoid the use of floating point arithmetic in the topology calculation algorithm, all of the above scores are scaled by a factor of 1000 and rounded to the nearest integer.

**Table 1.** Composition of transmembrane data set

| Protein class | Number in set |
|---|---|
| Prokaryotic | 98 |
| Eukaryotic | 65 |
| With signal peptides | 45 |
| Single-spanning TM segment | 52 |
| Multiple-spanning TM segments | 132 |
| Total | 184 |

In evaluating methods using the data set in Table 1, predictions were assessed using three criteria. Firstly, has the correct number of TM helices been predicted; secondly, has the correct topology been predicted; and thirdly, are the TM segments correctly located. To decide on the correct TM segment location, the author has followed the lead of other groups (e.g. Käll *et al.*, 2004), where a segment is deemed correctly located if there is at least a five-residue overlap with the observed segment in the testing set. For completeness, results based on a 10 residue overlap are also collated.

In evaluating MEMSAT3, the appropriate cross-validated neural networks were used in calculating performance statistics. When evaluating other methods, we were not able to cross-validate results, and the reported accuracies may therefore be slightly overestimated in those cases. All of the cross-validated neural nets are available, should users need to reproduce the fully cross-validated results described below.

In an attempt to improve the discrimination of transmembrane proteins from globular proteins, a second neural network was trained with the same number of inputs (399), seven hidden units and just a single output. In this case, the network was trained on the same set of 184 transmembrane proteins, but with the addition of 416 randomly chosen globular proteins (giving a total training set of 600 proteins) taken from a set of 2685 non-redundant chains from known protein structures. The single target output in this experiment indicated the presence of a TM segment at the centre of the sequence window. Signal peptides were included in the list of negative training cases. For cross-validation, each transmembrane protein was removed from the training set along with all homologues (as described above). Given the large surplus of globular protein test cases, cross-validation was carried out by evaluating on the full set of 2685 PDB entries, but excluding all of the 416 entries used for training (2269 test cases remaining). The complete test set therefore comprised 184 transmembrane proteins and 2269 globular proteins, giving a total of 2453 test cases.

## 3 RESULTS

Table 2 shows the overall results of applying MEMSAT3 to the test set of 184 proteins, with results for MEMSAT2 (Jones, 1998), along with more recent methods included for reference. MEMSAT3 results are fully cross-validated with all proteins homologous to the target removed from the neural net training set. For TMHMM, Phobius and PolyPhobius, the results were obtained from the respective Web servers and therefore are not cross-validated. In determining a correct prediction, both the predicted topology and the positions of prediction TM segments are assessed. Signal peptides are included in the target sequences but are not considered part of the observed

**Table 2.** Benchmark results for MEMSAT3 compared to other representative methods

| Method | Correct no. of TM segments | Correct topology | Correct topology and locations | Correct (overlap of 10) |
|---|---|---|---|---|
| MEMSAT2 | 125 (67.9%) | 108 (58.7%) | 104 (56.5%) | 98 (53.3%) |
| TMHMM | 131 (71.1%) | 116 (63.0%) | 114 (62.0%) | 108 (58.7%) |
| Phobius | 152 (82.6%) | 134 (72.8%) | 126 (68.4%) | 120 (65.2%) |
| PolyPhobius | 148 (80.4%) | 133 (72.2%) | 133 (72.2%) | 132 (71.7%) |
| MEMSAT3 | 156 (84.8%) | 150 (81.5%) | 147 (79.9%) | 141 (76.6%) |

topology (i.e. a signal peptide predicted to be a TM segment would be considered an incorrect topology prediction).

To allow comparison to previous benchmarking studies, the most relevant results are those in the third column, where a correct prediction must have not only the correct topology, but also an overlap of at least five residues between the predicted transmembrane segments and those which have been experimentally determined. In this case, MEMSAT3 is correct in 80% of cases (147 out of 184) compared to the next best method (PolyPhobius), which predicted 72.3% (133 out of 184) of cases correctly. Using a more stringent criterion of a 10-residue overlap gives a success rate of 76.6% for MEMSAT3, with PolyPhobius unchanged at 72.3%. In terms of the most basic criterion, predicting the correct number of TM segments irrespective of correct topology and location, both MEMSAT3 and Phobius achieve a similar level of accuracy.

In terms of over- and underpredictions, MEMSAT3 is fairly balanced. Over the set of 184 test cases, there are a total of 20 overpredicted TM segments and 24 missed (underpredicted) segments.

Table 3 shows a breakdown of the MEMSAT3 benchmarking results for difference subsets of the testing set. Compared to other methods, the performance of MEMSAT3 shows a stronger preference for prokaryotic sequences, with 87% (85 out of 98) of these proteins being correctly predicted. For eukaryotes, 71% (46 out of 65) of the targets are correctly predicted.

Despite the inclusion of limited signal peptide prediction in MEMSAT3, sequences without signal peptides are still overall more accurately predicted than sequences with them. For sequences without signal peptides present, 83% (123 out of 149) are correctly predicted. For sequences with signal peptides, MEMSAT3 is less effective than both Phobius and TMHMM, with only 63% (22 out of 35) correctly predicted.

The fact that MEMSAT3 does not show a marked difference between correctly predicting the number of TM segments and the correct topology clearly indicates that the topogenic signals picked up by the neural network are much stronger than the single amino acid scoring schemes employed in the other methods. To better quantify this, the neural network was evaluated in order to benchmark its ability to predict whether each residue in the test set is located in the cytoplasm or not. Each loop in the data set was evaluated by both the MEMSAT scoring tables and the MEMSAT3 neural network (with cross-validation) to see how many could be correctly classified as cytoplasmic or non-cytoplasmic. In the case of the MEMSAT3 neural network, 84% (805/964) of loops were correctly assigned compared to only 72% (690/964) for the original MEMSAT scoring tables. For the longest loops (length >50), the difference is more striking, with 85% (141/166) correctly assigned by MEMSAT3 compared to only 65% (106/166) by the MEMSAT scoring tables.

**Table 3.** Benchmark results for MEMSAT3 compared to other representative methods on different subsets of the benchmark data

| Subset | TMHMM | PolyPhobius | MEMSAT3 |
|---|---|---|---|
| Prokaryotic (98) | 63 (64.3%) | 73 (74.5%) | 85 (86.7%) |
| Eukaryotic (65) | 43 (66.2%) | 52 (80.0%) | 46 (70.8%) |
| With signal peptides (35) | 26 (74.3%) | 35 (100.0%) | 22 (62.9%) |
| Without Signal peptides (149) | 88 (59.1%) | 99 (66.4%) | 124 (83.2%) |
| Single-spanning (52) | 36 (69.2%) | 35 (67.3%) | 51 (98.1%) |
| Multiple-spanning (132) | 78 (59.1%) | 100 (75.8%) | 95 (72.0%) |
| All (184) | 114 (62.0%) | 133 (72.2%) | 147 (79.9%) |

Standard errors for all percentages were estimated with the bootstrapping procedure used by Chen et al. (2002) and were found to be below 0.5% in all cases.

Perhaps the most interesting aspect of the performance of MEMSAT3, however, is the very high success rate on single-spanning TM proteins in the data set. In the single-spanning case, 98% (51 out of 52) accuracy is achieved, compared to only 72% for multiple-spanning proteins.

### 3.1 Segment end point prediction accuracy

Figure 1 illustrates the reliability of MEMSAT3 in terms of TM segment end point prediction. For this, the analysis had to be limited to the 28 proteins in the data set with known 3-D structure and no large gaps in the structure. Across the total of 157 segments in this set, the mean error in predicting the start of each TM segment is 2.9 (standard deviation 2.4) and, for the ends, the mean error is 3.7 (standard deviation 2.7).

### 3.2 Overall reliability of topology predictions

Figure 2 illustrates the reliability of MEMSAT3 in terms of false-positive rate (i.e. fraction of predictions with incorrect

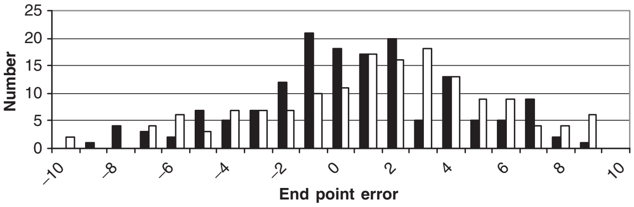

**Fig. 1.** End point prediction error distribution for MEMSAT3. Black bars are for the N-termini, white bars the C-termini. A negative error indicates that the predicted location starts or finishes earlier than the observed segment.

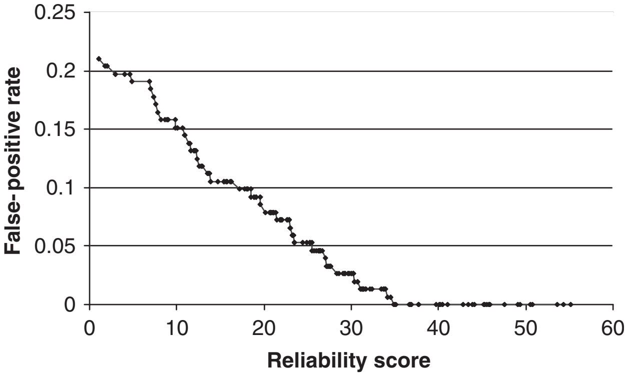

**Fig. 2.** False-positive rate for topology predictions plotted against reliability score threshold.

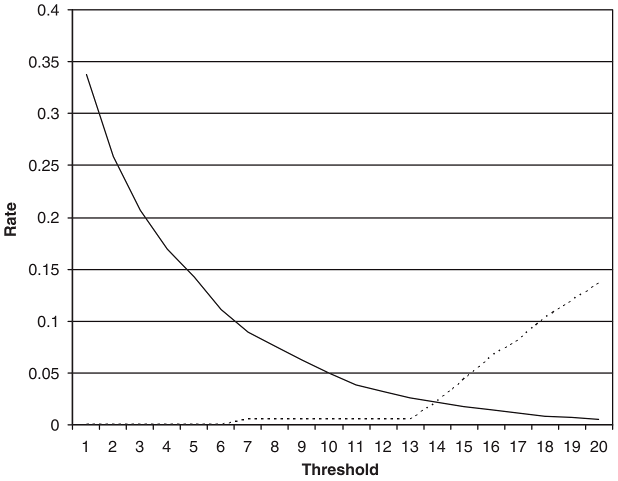

**Fig. 3.** False-positive rate (solid line) on identifying globular proteins and false-negative rate on transmembrane proteins (broken line) for the topology prediction algorithm is plotted against an overall score threshold. The threshold in this case is the number of residues predicted to be in a transmembrane segment.

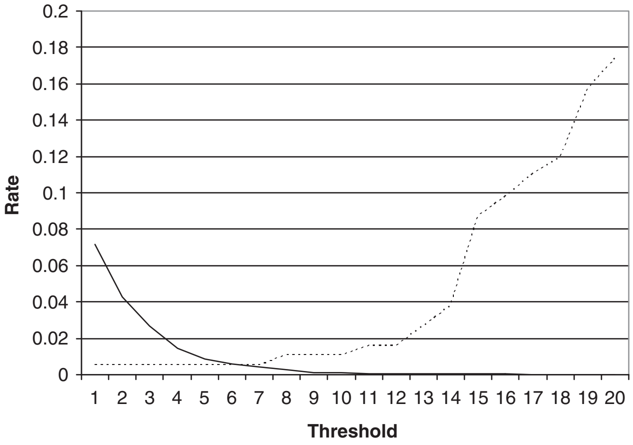

**Fig. 4.** False-positive rate (solid line) on globular proteins and false-negative rate on transmembrane proteins (broken line) for the neural network trained specifically to discriminate transmembrane from globular proteins. The threshold in this case is again the number of residues predicted to be in a transmembrane segment.

topology or mislocated TM segments) plotted against reliability score. The reliability score is simply the difference between the overall model score for the top prediction and the score for the next best scoring model (Melen *et al.*, 2003).

### 3.3 Discriminating globular and transmembrane proteins

Using the combined set of 2453 transmembrane and globular proteins, the topology prediction algorithm was evaluated in terms of its ability to discriminate transmembrane from globular proteins (Fig. 3).

As a discrimination criterion, a threshold was set on the number of residues predicted to be part of a transmembrane segment by the neural network. Increasing this threshold thus decreases the number of proteins predicted to be transmembrane. At the crossover point (threshold of 14), the method has a false-positive rate and false-negative rate of 2%.

To improve upon this, a second neural network was trained specifically to discriminate transmembrane from globular protein segments, as described earlier. In this case, the crossover point is found to give a much more satisfactory false-positive and false-negative rate of 0.5% (Fig. 4).

## 4 DISCUSSION AND CONCLUSIONS

The benchmarking results clearly show that MEMSAT3 is an effective method for transmembrane topology prediction, with higher overall prediction accuracy than other popular methods and the previous version of the method (MEMSAT2) based on statistical scoring tables. The major new source of information in MEMSAT3 is the evolutionary information manifest in the PSI-BLAST-derived sequence profiles, from which the neural network is able to determine much more reliable topogenic scores than could be obtained through single residue statistics. Although HMM-based methods (e.g. PolyPhobius) have attempted to make some use of multiple sequence information, the limitations of an HMM representation is such that multiple sequence information only ends up being used for finding a weighted average path through the underlying model states and thus only ends up improving prediction accuracy by a modest amount. What is missing from this simplified treatment of evolutionary information is the information that can be derived from an analysis of the conservation patterns themselves. For example, we know in globular proteins that conserved hydrophobic residues are likely to be buried in the core. From this, we can infer patterns of solvation that help identify strands and helices in secondary structure prediction. Similarly, for transmembrane proteins, we observe that highly variable hydrophobic positions in multiple-sequence alignments tend to indicate residues that are in contact with the lipid groups in the membrane (Donnelly *et al.*, 1993; Taylor *et al.*, 1994). Observing a pattern of conserved and variable hydrophobic positions with the correct periodicity of an alpha helix can help identify transmembrane helices and thus improve the overall accuracy of predictions. This information is not accessed when sequence profiles are used simply in a weighted average of match scores.

It is likely that the significant improvement in prediction accuracy for single-spanning TM proteins is a direct result of the use of sequence conservation. The membrane-spannning regions in monotopic TM proteins are very distinguishable due to the fact that all of the residues in the spanning regions are generally in contact with lipid, and thus these regions are overall highly mutable (yet still highly hydrophobic in character) compared to other TM segments and other hydrophobic regions in proteins. Hydrophobic regions in globular domains that are frequently predicted incorrectly as single-TM segments generally correspond either to buried secondary structural elements or to binding sites. In both of these cases, the sequences could be highly hydrophobic, but in both cases we would expect them to be evolutionarily highly conserved.

The use of a neural network to evaluate windows of a sequence profile in the scoring of protein topologies is clearly much more effective than applying *ad hoc* rules or single-residue statistical scoring tables. Individual residues in fact contribute very little topogenic information when taken in isolation. The positive-inside rule, for example, is a compositional rule where a greater concentration of positively charged residues are expected inside the cytoplasm, compared to outside. It is clearly of little sense to assume that every positively charged residue is found in the cytoplasm. In contrast, the neural network is able to detect useful topogenic information across the majority of the sequence and, in some cases, apparently strong localized topogenic features (i.e. short regions strongly predicted to be either inside or outside) are detected by the neural network in long loops (>50 residues) that seem (at least by eye) to have few distinguishing features. This hints at the fact that there could be as yet unknown topogenic sequence features buried in sequences that the neural network is able to pick up but which are not yet characterized in the literature. We hope to follow this up in a future study.

Although the overall prediction accuracy of MEMSAT3 was higher than other popular methods, its innate ability to discriminate transmembrane proteins from globular proteins was found to be slightly poorer than for other methods. Although overall accuracy is the most important issue for individual predictions (i.e. predictions made on proteins already suspected to be transmembrane proteins), the discrimination rate is important when a method is applied to whole-genome annotation. A simple but effective compromise solution was found whereby a second neural network was employed specifically to decide whether the target protein is likely to contain transmembrane segments or not. The key difference for this second network was to train it on a mixture of transmembrane and globular proteins. In practice, this network is employed as a pre-filter to decide whether or not the MEMSAT3 algorithm should be applied.

Some recent X-ray structures of alpha-helical integral membrane proteins (e.g. the voltage-gated potassium channels) have shown some unexpected features, such as very short helices that do not span the bilayer or very long membrane-spanning helices. Long helices prove not to be a particular problem, as MEMSAT3 easily allows longer helices to be generated, but helices that only traverse the bilayer halfway and then return to the same membrane face violate the assumption that the topology is a simple meander from one bilayer face to the other. In the case of subunits of the potassium channel, for example, MEMSAT correctly predicts three TM segments (24–42; 50–66; 78–95), but predicts that only the N-terminus is cytoplasmic, whereas in reality both termini are located in the cytoplasm. Interestingly, the raw neural network output clearly indicates a preference for the C-terminus to be cytoplasmic and the second best scoring topology (with a score very close to the optimum) skips the partial TM helix and allows both termini to be cytoplasmic. It is possible that a rule-based approach might be used in future to help identify these situations.

One major area where MEMSAT3 can clearly be improved further is in the handling of signal peptides. The Phobius method has already demonstrated how powerful the use of signal peptide prediction can be in helping to improve transmembrane topology prediction, and outperforms MEMSAT3 on proteins with signal peptides. Although MEMSAT3 already includes a simple signal peptide predictor, the relative contribution of the signal peptide output from the neural network was surprisingly small. This is likely due not just to the simple algorithm but also the small sample size of signal peptides in the training set. Clearly there also needs to be some consideration as to whether the origin of the target sequence is prokaryotic or eukaryotic in the prediction of signal peptides and the lack of this information also no doubt contributes to the poor performance on signal peptides. It is expected in future developments of the MEMSAT method that the addition of an explicit signal peptide prediction stage will give significant benefit in prediction accuracy.

## ACKNOWLEDGEMENTS

This work was partly supported by the Biosapiens Network of Excellence, which is funded by the European Commission within its FP6 Programme, under the thematic area ‘Life sciences, genomics and biotechnology for health’, contract number LSHG-CT-2003-503265.

*Conflict of Interest*: none declared.

## REFERENCES

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

Amico,M. *et al.* (2006) *Nucl. Acids Res.*, **34**, W169–W172.

Bowie,J.U. (2005) Solving the membrane protein folding problem. *Nature*, **438**, 581–589.

Berman,H.M. *et al.* (2000). The Protein Data Bank. *Nucl. Acid Res.*, **28**, 235–242.

Bowie,J.U. (2005) Solving the membrane protein folding problem. *Nature*, **438**, 581–589.

Chen,C.P. *et al.* (2002) Transmembrane helix predictions revisited. *Protein Sci.*, **11**, 2774–2791.

Donnelly,D. *et al.* (1993) Modeling alpha-helical transmembrane domains: the calculation and use of substitution tables for lipid facing residues. *Protein Sci.*, **2**, 55–70.

Fleishman,S.J. and Ben-Tal,N. (2006) Progress in structure prediction of α-helical membrane proteins. *Curr. Opin. Struct. Biol.*, **16**, 496–504.

Melen,K. *et al.* (2003) Reliability measures for membrane protein topology prediction algorithms. *J. Mol. Biol.*, **327**, 735–744.

Jones,D.T. *et al.* (1994) A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry*, **33**, 3038–3049.

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

Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.

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

Käll,L. *et al.* (2005) An HMM posterior decoder for sequence feature prediction that includes homology information. *Bioinformatics*, **21** (Suppl 1), i251–i257.

Kyte,J. and Doolittle,R.F. (1982) A simple method for displaying the hydropathic character of proteins. *J. Mol. Biol.*, **157**, 105–132.

Martelli,P.L. *et al.* (2003) An ENSEMBLE machine learning approach for the prediction of all-alpha membrane proteins. *Bioinformatics*, **19**, i205–i211.

Moller,S. *et al.* (2001) Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics*, **17**, 646–653.

Nilsson,J. *et al*. (2002) Prediction of partial membrane protein topologies using a consensus approach. *Prot. Sci*., **11**, 2974–2980.

Rost,B. *et al*. (1996) Topology prediction for helical transmembrane proteins at 86% accuracy. *Prot. Sci*., **4**, 521–533.

Sonnhammer,E.L.L. *et al*. (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. *Intell. Syst. Mol. Biol*., **6**, 175–182.

Tusnady,G.E. and Simon,I. (1998) Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *J. Mol. Bio*., **283**, 489–506.

Taylor,W.R. *et al*. (1994) A method for alpha-helical integral membrane protein fold prediction. *Proteins*, **18**, 281–294.

von Heijne,G. (1992) Membrane protein structure prediction. *J. Mol. Biol*., **255**, 487–494.

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

White,S.H. and Wimley,W.C. (1999) Membrane protein folding and stability: physical principles. *Ann. Rev. Biophys. Struct*., **28**, 319–365.

White,S.H. (2001) How membranes shape protein structure. *J. Biol. Chem*., **31**, 32395–32398.

Yarov-Yarovoy,V. *et al*. (2006) Multipass membrane protein structure prediction using Rosetta. *Proteins*, **62**, 1010–1025.