Procedings of ISMB 6, 1998, pages 175-182

A hidden Markov model for predicting transmembrane helices in protein sequences

Erik L.L. Sonnhammer
National Center for
Biotechnology Information
Building 38A, Room 8N805, NLM/NIH
Bethesda, Maryland 20894, USA
Email: esr@ncbi.nlm.nih.gov

Gunnar von Heijne
Department of Biochemistry
Arrhenius Laboratory
Stockholm University
S-106 91 Stockholm, Sweden
Email: gunnar@biokemi.su.se

Anders Krogh
Center for Biological Sequence Analysis
Technical University of Denmark
Building 208, 2800 Lyngby, Denmark
E-mail: krogh@cbs.dtu.dk

Abstract

A novel method to model and predict the location and orientation of alpha helices in membrane- spanning proteins is presented. It is based on a hidden Markov model (HMM) with an architecture that corresponds closely to the biological system. The model is cyclic with 7 types of states for helix core, helix caps on either side, loop on the cytoplasmic side, two loops for the non-cytoplasmic side, and a globular domain state in the middle of each loop. The two loop paths on the non-cytoplasmic side are used to model short and long loops separately, which corresponds biologically to the two known different membrane insertions mechanisms. The close mapping between the biological and computational states allows us to infer which parts of the model architecture are important to capture the information that encodes the membrane topology, and to gain a better understanding of the mechanisms and constraints involved. Models were estimated both by maximum likelihood and a discriminative method, and a method for reassignment of the membrane helix boundaries were developed. In a cross validated test on single sequences, our transmembrane HMM, TMHMM, correctly predicts the entire topology for 77% of the sequences in a standard dataset of 83 proteins with known topology. The same accuracy was achieved on a larger dataset of 160 proteins. These results compare favourably with existing methods.

Introduction

Prediction of membrane-spanning alpha helices in proteins is a frequent sequence analysis objective. A large portion of the proteins in a genome encode integral membrane proteins (Himmelreich *et al*. 1996; Frishman & Mewes 1997; Wallin & von Heijne 1998). Knowledge of the presence and exact location of the transmembrane helices is important for functional annotation and to direct functional analysis.

Transmembrane helices are substantially easier to predict than helices in globular domains. Predicting 95% of the transmembrane helices in the 'correct' location is not unusual (Cserzo *et al*. 1997; Rost *et al*. 1995). By 'correct' is meant that the prediction overlaps the true location. The reason for this high accuracy is that most transmembrane alpha helices are encoded by an unusually long stretch of hydrophobic residues. This compositional bias is imposed by the constraint that residues buried in lipid membranes must be suitable for hydrophobic interactions with the lipids. The hydrophobic signal is so strong that a straightforward approach of calculating a propensity scale for residues in transmembrane helices and applying a sliding window with a cutoff already performs quite well.

In addition to knowing the location of a transmembrane helix, knowledge of its orientation, i.e. whether it runs inwards or outwards, is also important for making functional inferences for different parts of the sequence. The orientations of the transmembrane helices give the overall *topology* of the protein.

It is known that the positively charged residues arginine and lysine play a major role in determining the orientation as they are mainly found in non-transmembrane parts of the protein ('loops') on the cytoplasmic side (von Heijne 1986; Jones, Taylor, & Thornton 1994; Persson & Argos 1994; Wallin & von Heijne 1998), often referred to as the 'positive-inside rule'. Since the rule also applies to proteins in the membrane of intracellular organelles (Gavel *et al*. 1991; Gavel & von Heijne 1992), we shall use the terms 'cytoplasmic' and 'non-cytoplasmic' for the two sides of a membrane.

The difference in amino acid usage between cytoplasmic and non-cytoplasmic loops can be exploited to improve the prediction of transmembrane helices by validating potential transmembrane helices by the charge bias they would produce (von Heijne 1992). Despite this relatively consistent *topogenic* signal, correct prediction of the location and orientation of all transmembrane segments has proved to be a difficult problem. On a reasonably large dataset of single sequences, a topology accuracy of 77% has been reported (Jones, Taylor, & Thornton 1994), and aided with multiple alignments 86% (Rost, Fariselli, & Casadio 1996). The difficulty in predicting the topology seems to be partly caused by the fact that the positive-inside rule can be blurred by globular domains in loops on the non-cytoplasmic side that contain a substantial number of positively charged residues.

It has been shown that positively charged residues in short loops guide the orientation of helices by preventing translocation across the membrane (von Heijne 1994). However, long loops containing positively charged residues do not necessarily arrest the translocation, and may be transferred across the membrane by a special mechanism. This has been shown in bacteria (Andersson & von Heijne 1994).

The first method to predict the complete topology of transmembrane proteins, TopPred (von Heijne 1992), applies two empirical hydrophobicity cutoffs to the output of a sliding trapezoid window in order to compile a list of certain and putative transmembrane helices. The combination of putative helices that produces the strongest enrichment of positively charged residues in loops on the cytoplasmic side is selected as the best prediction. Loops that are longer than 70 residues are ignored.

A potential drawback of TopPred and other methods that depend on fixed hydrophobicity thresholds for considering a segment as a transmembrane helix is that some helices may be missed that fall just under the threshold. This is not unusual in proteins with many membrane-spanning helices that form a bundle in which non-hydrophobic residues may make contacts between helices.

A number of approaches have been explored to improve prediction accuracy. Memsat (Jones, Taylor, & Thornton 1994) performs a constrained dynamic programming that incorporates hydrophobicity and topogenic signals to find the optimal location and orientation of a given number of transmembrane helices. It uses separate propensity scales for residues in the head and the tail region of the membrane (allowed to be 4 and 17-25 residues respectively). The highest scoring number of transmembrane helices is selected as the best prediction. PHDhtm (Rost, Fariselli, & Casadio 1996) uses a neural network to predict transmembrane segments. A second postprocessing step is applied to find the best consistent combination of these segments that maximises the positive residue content on the cytoplasmic side. PHDhtm automatically generates a multiple alignment of the query and its homologues, and performs the prediction on the multiple alignment, which improves the accuracy significantly. TMAP (Persson & Argos 1997) scans either single sequences or multiple alignments for peaks in propensity curves for the head and tail regions, and uses the frequency biases of twelve kinds of amino acids to predict the topology.

In this paper, we introduce the probabilistic framework of the hidden Markov model (HMM) to transmembrane helix prediction. Hidden Markov models have been used successfully in computational biology to model e.g. the statistical structure of genomes (Churchill 1992), protein families (Krogh *et al.* 1994; Eddy 1996) and gene structure (Kulp *et al.* 1996; Krogh 1997). The basic principle is to define a set of states, each corresponding to a region or specific site in the proteins being modelled. In the simplest case, a model for a transmembrane protein may consist of three states: one for inside loops, one for transmembrane regions, and one for outside loops. Each state has an associated probability distribution over the 20 amino acids characterising the variability of amino acids in the region it models. The states are connected to each other in a biologically reasonable way, so for instance the state for inside loop is connected to itself, because loops may be longer than 1, and to the transmembrane helix state, because after an inside loop a helix begins. A 'transition probability' is associated with each transition. The amino acid probabilities and the transition probabilities are learned by a standard inference techniques that computes the maximum posterior probabilities given a prior and the observed frequencies.

By defining states for transmembrane helix residues and other states for residues in loops, residues on either side of the membrane, and connecting them in a cycle, we can produce a model that in architecture closely resembles the biological system we are modelling. If the model parameters are tuned to capture the biological reality, the path of a protein sequence through the states with the highest probability should be able to predict the true topology. Since the HMM method does not employ any fixed empirical cutoffs or rules, and since the optimal path through the HMM is found in a single step, it should be more flexible to handle cases where several signals need to be combined to find the true topology. For instance, a segment that normally would not be considered a transmembrane helix due to poor hydrophobicity may still be predicted if the surrounding topogenic signals strongly support it. Such helices are fairly common in multi-spanning proteins where the transmembrane helices have hydrophilic interactions with each other.

We believe that apart from achieving high prediction accuracy, the fact that the model corresponds well to the biology is also very important. We have therefore only sampled architectures that make biological sense. By varying the model we can explore what architectural features are most important for successful prediction, and thus learn biologically meaningful rules. For instance, we can easily explore what the optimal head region length is, and whether separate paths for long or short loops is better than one.

This paper describes the basic principles of TMHMM and presents prediction results on single sequences. We have not extended it to work on multiple alignments here, partly because that would make evaluation of the algorithm *per se* harder.

Methods

Architecture of the HMM

The basic architecture of TMHMM is shown in Figure 1. There are three main locations of a residue: in the transmembrane helix core (in the hydrophobic tail region of the membrane), in the transmembrane helix caps (in head region of the membrane), and in loops. Due to the different residue distributions on the different sides however, we use seven different states: one for the helix core, two for caps on either side, one for loops on the cytoplasmic side, one each for short and long loops on the non-cytoplasmic side, and one for 'globular domains' in the middle of each loop. The amino acid emission probabilities of all states of the same type are 'tied' to each other, i.e. they are estimated collectively.

The transmembrane helix is modelled by two cap regions

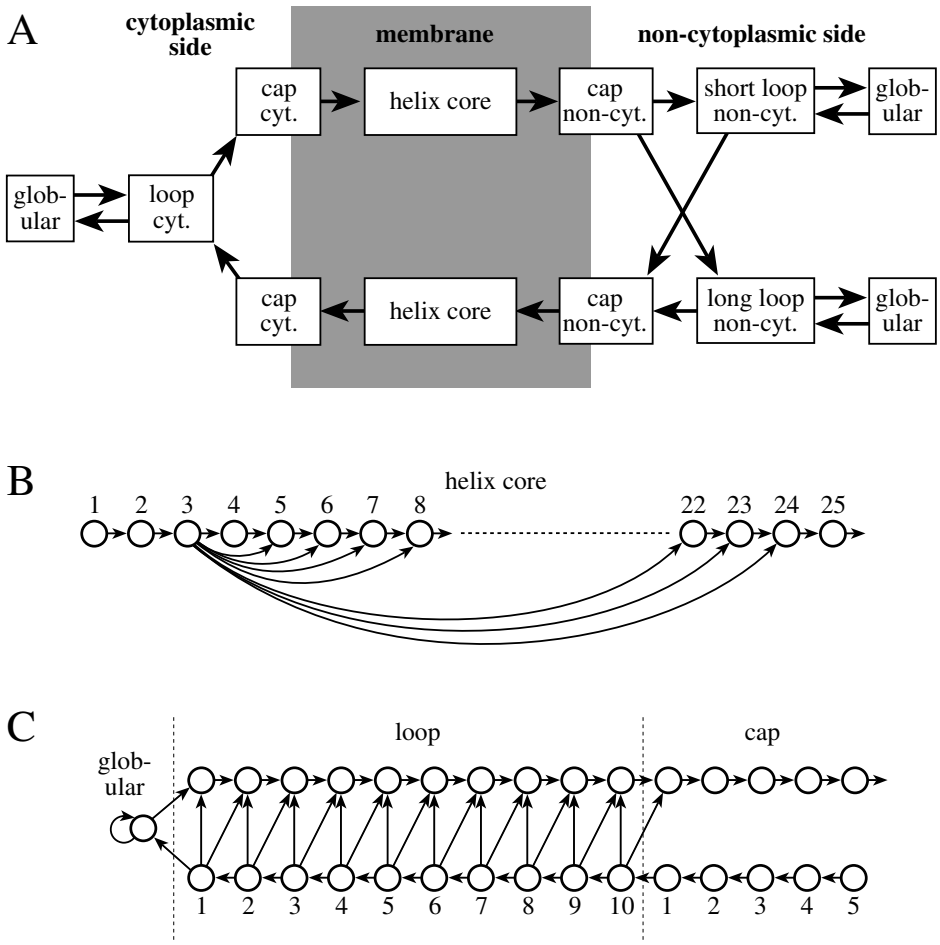

Figure 1: The structure of the model used in TMHMM. A) The overall layout of the model. Each box corresponds to one or more states. Parts of the model with the same text are tied, i.e. their parameters are the same. Cyt. means the cytoplasmic side of the membrane and non-cyt. the other side. B) The state diagram for the parts of the model denoted helix core in A. From the last cap state there is a transition to core state number 1. The first three and the last two core states have to be traversed, but all the other core states can be bypassed. This models core regions of lengths from 5 to 25 residues. All core states have tied amino acid probabilities. C) The state structure of globular, loop, and cap regions. In each of the three regions the amino acid probabilities are tied. The three different loop regions are all modelled like this, but they have different parameters in some regions.

of 5 residues each, surrounding a core region of variable length 5-25 residues. This allows for helices 15-35 residues long. The flanking cap regions have their own amino acid distributions (one for the cytoplasmic side and one for the non-cytoplasmic side), but they are labelled the same way as helical core residues both during training and prediction. Although the model contains two sets of transmembrane states to model paths going inwards and outwards, all their parameters are mirrored and tied to each other. There are 21 variable transition probabilities to model the length distribution of the helical core, and because they sum to one it corresponds to 20 free parameters.

The loops between the helices are modelled by modules that contain 2×10 states in a ladder configuration, and one self-looping state. The idea is that the 10 first states should contain most of the topogenic signals (bias in amino acid usage) while larger, globular domains are modelled in a simple way by the single self-looping state, which has a neutral amino acid distribution.

Long loops on the non-cytoplasmic side that contain globular domains appear to have different properties than short loops. They do not consistently exhibit the sparseness in positively charged residues observed in short non-cytoplasmic loops. We therefore model non-cytoplasmic loops by two different pathways. The HMM thus contains two parallel loop modules on that side. During training, loops on the non-cytoplasmic side longer than 100 residues are given a special label that directs them to the appropriate module. Each loop module contains 21 free transition probabilities.

The total number of free parameters in the entire model is thus $7 \times 19 + 20 + 3 \times 21 = 216$. This should be compared to neural networks, that usually contain tens of thousands (Rost, Fariselli, & Casadio 1996).

Training the HMM

The estimation of the model parameters proceeded in stages.

In the first stage, the model was estimated by the Baum-Welch reestimation procedure, which is the standard method for maximum likelihood estimation of HMMs, see for instance (Rabiner 1989; Krogh *et al.* 1994; Durbin *et al.* 1998). However, the reestimation was done from labeled sequences as described in (Krogh 1994; 1997), because that allows one to use 'soft boundaries' of transmembrane helices (see below). To avoid local maxima of the likelihood, a simulated annealing scheme was used that allows unfavourable models to be sampled with some probability. This was done by adding noise to the model parameters, and then decreasing the level of this noise by 5% per iteration, see e.g. (Hughey & Krogh 1996). Compared to HMMs of e.g. protein families, TMHMM contains very few parameters and training is therefore usually quick and reproducible. After convergence, the most likely path of a query sequence is calculated with the Viterbi algorithm (Rabiner 1989; Krogh *et al.* 1994; Durbin *et al.* 1998).

The precise end of a transmembrane helix is usually only approximately known. Many transmembrane segments are therefore annotated as a segment of somewhat arbitrary length, often 21 residues. A technique to accommodate misplaced borders in the training data is to 'dilute' the labels by unlabelling a few states at each label boundary (i.e. between loop and membrane) in order to allow some freedom in choosing the state. We used three unlabelled residues on each side of a border since this appeared to give the best results (although there is very little difference between 2 and 3). It means that during training, the exact helix boundaries are put where they fit the current model the best.[1]

In the second stage of model estimation, the first model was used to *relabel* the data. The labeling was diluted as described above, but this time 5 residues were unlabeled to each side of a helix boundary (but such that at least one label remained in both the left and right region), see Figure 2. Then the first model was used to predict transmembrane segments consistent with the remaining labels. More formally, this means that the most probable path through the model is found, subject to the constraint that the prediction conforms with the diluted labels. This gives a new labeling consistent with the important structure of the protein, but with the exact boundaries moved such that it fits the model better, see Figure 2.

The second model was trained from the relabeled sequences with no further unlabeling. In this stage there was no need for simulated annealing, because there was no uncertainty about boundaries.

In both of the first stages the models were weakly regularised. This was done by adding 'pseudocounts' to the estimated counts used in the re-estimation procedure. When regularising this way, the estimation procedure can be seen as maximum aposteriori estimation instead of maximum likelihood, see e.g. (Krogh *et al.* 1994; Durbin *et al.* 1998). The distribution of amino acids were found for transmembrane helices in the training set and these numbers were used as pseudocounts in the part of the model for transmembrane segments. Similarly for the loop regions on both the cytoplasmic and the non-cytoplasmic side. Only the model of 'globular domains' was strongly regularised by adding pseudocounts of 1000 times a standard neutral background amino acid distribution. This was done in order to prevent the model from learning some skewed distribution of amino acids in globular domains due to biasses in the training data. The size of these psudocounts is arbitrary, but changing it has almost no effect on performance. In fact, one can fix the distribution in the globular state (corresponding to extremely large pseudocounts) without changing performance significantly. The parameters for the initial models were obtained by normalising the pseudocounts appropriately (before adding noise).

In the third stage the second model was trained further by a method for 'discriminative' training (Krogh 1994; 1997). This training method aims at maximising the probability of the correct prediction rather than optimising the model of the protein. Discriminative training is more prone to over-fitting, and therefore this model was regularised heavily by the maximum likelihood model. The size of the

[1] In the Baum-Welch algorithm, the reestimated model is actually a result of summing over all possible boundaries consistent with the diluted labeling.

```
       TAL6_HUMAN

           MCYGKCARCIGHSLVGLALLCIAANILLYFPNGETKYASENHLSRFVWFFSGIVGGGLLMLLPAFVFIGL
Correct    iiiiiiiiiMMMMMMMMMMMMMMMMMMMMoooooooooooooooMMMMMMMMMMMMMMMMMMMMMMMM
Unlabeled  iiii.........MMMMMMMMMM.........ooooo.........MMMMMMMMMMMMMMM.....
Relabeled  iiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMMoooooooooooooooMMMMMMMMMMMMMMMMMMMMM

           EQDDCCGCCGHENCGKRCAMLSSVLAALIGIAGSGYCVIVAALGLAEGPLCLDSLGQWNYTFASTEGQYL
Correct    iiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMMMMMMMMooooooooooooooooooooooooo
Unlabeled  .....iiiiiiii..........MMMMMMMMMMMMMMMMMM.........ooooooooooooooooooo
Relabeled  iiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMMMMooooooooooooooooooooooooooo

           LDTSTWSECTEPKHIVEWNVSLFSILLALGGIEFILCLIQVINGVLGGICGFCCSHQQQYDC
Correct    ooooooooooooooooooooooMMMMMMMMMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiii
Unlabeled  oooooooooooooooo..........MMMMMMMMMMMMMMMMMMMMM.........iiiii
Relabeled  ooooooooooooooooooooooooMMMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiii
```

Figure 2: First line after the sequence shows the labeling of the sequence as obtained from Swissprot. Cytoplasmic is labeled by 'i' (inside), transmembrane helices by 'M', and non cytoplasmic by 'o' (outside). The second line shows the labeling after the boundaries have been unlabeled by 5 residues to each side. Unlabeled positions are indicated with a '.'. Finally a prediction is shown, which was obtained by forcing the prediction to conform with the labels, but let the program choose freely in the unlabeled regions.

pseudocounts were a constant times the probabilities found in the second stage. For the small set of proteins the constant was 400 and for the large database it was 600. The results are not very sensitive to this parameter, which was chosen after a few experiments. For the predictions, the one-best algorithm (Schwarz & Chow 1990) was used, because it is known to work better with discriminative training (Krogh 1997).

Datasets

We used two datasets. For comparison with other methods, we used the set of 83 proteins originally compiled by (Jones, Taylor, & Thornton 1994), as provided by Rost et al. (1996). It is marginally different from the dataset used by Jones et al., which we were unable to completely reconstruct. It consists of 38 multi-spanning and 45 single-spanning proteins whose topologies have been experimentally determined. We refer to this set as set 1. We have also compiled a larger set of 160 proteins, most of which have experimental topology data, which we refer to as set 2. It contains 108 multi-spanning and 52 single-spanning proteins. Most of the experiments to analyse the effect of different model architectures and training procedures were performed on set 2.

It should be noted that nearly all proteins with an 'experimentally determined' topology have been analysed with biochemical and genetic methods that are not always reliable (Traxler, Boyd, & Beckwith 1993). Only a very small number of membrane protein structures have been determined at an 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 incorrect data as much as possible, we did not include proteins in set 2 for which different experiments had yielded conflicting topologies and where it was not obvious which topology was closer to the truth.

The accuracy of the HMM was tested by 10-fold cross validation. For this the datasets were divided into 10 sets of about equal size so that no sequence was more than 25% identical to a sequence in another set (in a Needleman-Wunsch alignment). The HMM was trained on nine sets and predictions were made on the remaining set. This was repeated for all 10 different partitions of test and training sets, and in the end the accuracy was calculated on the predictions obtained on the test sets.

Results

Accuracy of the best model

The accuracy of the best model architecture is listed in Table 1, along with the results we obtained with Memsat. Both were applied to single sequences only. For comparison we also give the results of PHDhtm reported in (Rost, Fariselli, & Casadio 1996) for the set of 83 sequences (using multiple alignments). A predicted helix is counted as correct if it overlaps by at least 5 residues with a true helix (using an overlap of 1 instead of 5 only affect results marginally).

In the comparison between TMHMM and Memsat, only the TMHMM results were properly cross validated. The authors of Memsat claim to have obtained an accuracy of 77.1% correct topologies in a cross validated experiment on a dataset nearly identical to set 1, although we get a slightly lower number even without cross validation. When tested on the 77 sequences in set 2 that are not in set 1, Memsat had an accuracy of only 56%.

The architecture of transmembrane proteins

Many facts about the architecture of transmembrane proteins have been estimated from biochemical evidence and knowledge about the physicochemical properties of membranes

Method	Training set size	Stage of training	Correct topology	Correct locations	Single TM sensitivity	Single TM specificity
TMHMM	83	1	60 (72.3%)	62 (74.7%)	95.6%	96.4%
	83	2	63 (75.9%)	65 (78.3%)	96.2%	96.5%
	83	3	**64 (77.1%)**	**69 (83.1%)**	**96.2%**	**97.6%**
MEMSAT	83		63 (75.9%)	67 (80.7%)	96.8%	94.6%
PHDhtm	83		(85.5%)	(88.0%)	98.8%	95.2%
TMHMM	160	1	106 (66.3%)	122 (76.3%)	95.4%	97.1%
	160	2	120 (75.0%)	133 (83.1%)	96.8%	97.5%
	160	3	**123 (76.9%)**	**134 (83.8%)**	**97.1%**	**97.7%**
MEMSAT	160		108 (67.5%)	118 (73.8%)	93.3%	95.6%

Table 1: Transmembrane topology prediction accuracy of TMHMM, MEMSAT, and PHDhtm. The TMHMM values were measured in cross validated experiments while MEMSAT was run on the complete training set with default propensities. All predictions for TMHMM and MEMSAT were made using single sequences only, whereas PHDhtm uses multiple alignments. Results for PDHhtm are from (Rost, Fariselli, & Casadio 1996). The Stages of training are explained in the text.
Correct topology: proteins for which all transmembrane segments and their orientations are correctly predicted. *Correct locations*: proteins for which all transmembrane segments are correctly predicted, regardless of their orientation. *Single TM sensitivity*: correctly predicted segments/true segments. *Single TM specificity*: correctly predicted segments/total predicted segments.

and amino acids. Here we have the opportunity to derive some rules of transmembrane helices by experimenting with various HMM specifications and observing which architecture performs best in a cross validated test.

Length of helix cap region. The part of the helix that lies in the head region of the lipid bilayer contains many polar and charged residues that make contact to the phosphate groups of the lipids. The length of this region is often arbitrarily taken as being four residues (Jones, Taylor, & Thornton 1994). In TMHMM, the cap is modelled with a number of states with a separate amino acid distribution, flanking the central helix region (See Figure 1). We experimented with cap lengths between 0 and 7 residues, which is the maximum length if we allow helices down to 15 residues. We found that with less than 4 residues the accuracy drops significantly, while caps of 4-7 residues gave the same result.

Helix caps on different sides of the membrane are treated separately. Some prediction methods assume that the cap regions on both sides of the membrane should have the same amino acid distributions. However, we observed an accuracy reduction of up to 10% of proteins with all helices predicted correctly, and of up to 5% of correctly predicted topologies from tying the cap distributions together.

Loop architecture. Positively charged residues are predominantly found in loops on the cytoplasmic side. However, long globular domains with positively charged residues are equally often found on both sides of the membrane (Sipos & von Heijne 1993). Some prediction algorithms take this into account, for instance by ignoring any loop longer than 70 residues (von Heijne 1992), or by only considering the 25 residues nearest predicted transmembrane helices (Rost, Fariselli, & Casadio 1996).

In TMHMM, loops are modelled with two types of states: a number of states for the helix- flanking topogenic sequence, and a single self-looping state for larger globular domains (see Figure 1). The topogenic states are arranged in a ladder that captures the length distribution of short topogenic loops. Since it is believed that there may be two different mechanisms for translocating short and long 'globular domain' loops across the membrane, we tried using two alternative paths on the non-cytoplasmic side in TMHMM. We found that this increased the accuracy by 6-14%, when training the 'short' path on loops shorter than 100 residues and the 'large globular domain' path on all longer loops. There does however not seem to be an advantage in having two alternative loop paths on the cytoplasmic side; this reduced the accuracy by 2-11%. The highest accuracy was observed at loop ladder lengths between 2x10 and 2x15.

Availability

A model specification of TMHMM including all optimised parameters is available from

http://www.cbs.dtu.dk/krogh/TMHMM/

The used datasets, including topology-labels and our divisions into cross validation subsets are also provided. We plan to make a prediction server available via the World Wide Web. A binary UNIX program for finding the most likely topology of a query sequence can also be retrieved upon request. Please see the WWW page for details.

Discussion

Our HMM-based method embodies many conceptual and methodological aspects of previous methods. The main virtues are that the model architecture maps closely to the biological system, and that everything is done in the probabilistic framework of HMMs, that is, we do not have to develop a specialised dynamic programming algorithm or post processing method.

The accuracy of the TMHMM is high compared to MEMSAT, particularly on dataset 2. We were unable to compare TMHMM with PHDtmh on dataset 2, but could compare to published figures for dataset 1. Given that the results

of TMHMM were based on single sequences, we were surprised to see that it obtained about the same single TM accuracy as did PHDthm using multiple alignments. For overall topology, the accuracy of TMHMM is however not quite as high as with PHDtmh on this dataset.

At present, TMHMM only reports the most likely path of a sequence through the model. In many cases, however, it is desirable to report a number of top-scoring matches, particularly if they have similar scores. We plan to add this feature to TMHMM in the future. We also plan to make use of multiple alignments to increase the accuracy further.

We have here worked with a mix of transmembrane proteins from different sources and of different types, but have treated them as a unified set in order to find general principles. There is however evidence for differences in the membrane insertion mechanism between prokaryotic and eukaryotic proteins (Gafvelin *et al*. 1997). Preliminary experiments did not suggest that splitting the data up into these groups improved accuracy. This may partly be due to the fact that the subsets became too small for efficient training. It has also been suggested that single-spanning transmembrane proteins have distinct properties from multi-spanning proteins (Jones, Taylor, & Thornton 1994). In fact, MEMSAT uses different propensity tables for these types of models. It would in principle be possible to adapt our HMM to choose between two such specialised models. It is not clear whether this would be as much of an advantage to TMHMM as it is to MEMSAT, however. Since MEMSAT always predicts at least one helix, it may be needed for increased stringency. The most likely path through the HMM (i.e. the prediction) on the other hand, may contain no transmembrane helices at all.

Acknowledgements

We thank Henrik Nielsen, Arne Elofsson and Erik Wallin for helpful discussions. This work was supported by the Danish National Research Foundation and the Swedish Natural Sciences Research Council.

References

Andersson, H., and von Heijne, G. 1994. Positively charged residues influence the degree of SecA dependence in protein translocation across the E. coli inner membrane. *FEBS Lett*. 347(2-3):169–172.

Churchill, G. A. 1992. Hidden markov chains and the analysis of genome structure. *Computers and Chemistry* 16(2):107–115.

Cserzo, M.; Wallin, E.; Simon, I.; von Heijne, G.; and Elofsson, A. 1997. Prediction of transmembrane alpha-helices in prokaryotic membrane proteins: the dense alignment surface method. *Protein Eng*. 10(6):673–676.

Durbin, R. M.; Eddy, S. R.; Krogh, A.; and Mitchison, G. 1998. *Biological Sequence Analysis*. Cambridge University Press. To appear.

Eddy, S. R. 1996. Hidden Markov models. *Current Opinion in Structural Biology* 6:361–365.

Frishman, D., and Mewes, H. W. 1997. Protein structural classes in five complete genomes. *Nat. Struct. Biol.* 4(8):626–628.

Gafvelin, G.; Sakaguchi, M.; Andersson, H.; and von Heijne, G. 1997. Topological rules for membrane protein assembly in eukaryotic cells. *J. Biol. Chem*. 272(10):6119–6127.

Gavel, Y., and von Heijne, G. 1992. The distribution of charged amino acids in mitochondrial inner membrane proteins suggests different modes of membrane integration for nuclearly and mitochondrially encoded proteins. *Eur J Biochem* 205:1207–1215.

Gavel, Y.; Steppuhn, J.; Herrmann, R.; and von Heijne, G. 1991. The positive-inside rule applies to thylakoid membrane proteins. *FEBS Lett*. 282:41–46.

Himmelreich, R.; Hilbert, H.; Plagens, H.; Pirkl, E.; Li, B. C.; and Herrmann, R. 1996. Complete sequence analysis of the genome of the bacterium Mycoplasma pneumoniae. *Nucleic Acids Research* 24(22):4420–4449.

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

Jones, D. T.; Taylor, W. R.; and Thornton, J. M. 1994. A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry* 33(10):3038–3049.

Krogh, A.; Brown, M.; Mian, I. S.; Sjölander, K.; and Haussler, D. 1994. Hidden Markov models in computational biology: Applications to protein modeling. *Journal of Molecular Biology* 235:1501–1531.

Krogh, A. 1994. Hidden Markov models for labeled sequences. In *Proceedings of the 12th IAPR International Conference on Pattern Recognition*, 140–144. Los Alamitos, California: IEEE Computer Society Press.

Krogh, A. 1997. Two methods for improving performance of a HMM and their application for gene finding. In Gaasterland, T.; Karp, P.; Karplus, K.; Ouzounis, C.; Sander, C.; and Valencia, A., eds., *Proc. of Fifth Int. Conf. on Intelligent Systems for Molecular Biology*, 179–186. Menlo Park, CA: AAAI Press.

Kulp, D.; Haussler, D.; Reese, M. G.; and Eeckman, F. H. 1996. A generalized hidden Markov model for the recognition of human genes in DNA. In States, D.; Agarwal, P.; Gaasterland, T.; Hunter, L.; and Smith, R., eds., *Proc. Conf. on Intelligent Systems in Molecular Biology*, 134–142. Menlo Park, CA: AAAI Press.

Persson, B., and Argos, P. 1994. Prediction of transmembrane segments in proteins utilising multiple sequence alignments. *Journal of Molecular Biology* 237(2):182–192.

Persson, B., and Argos, P. 1997. Prediction of membrane protein topology utilizing multiple sequence alignments. *J. Protein Chem*. 16(5):453–457.

Rabiner, L. R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* 77(2):257–286.

Rost, B.; Casadio, R.; Fariselli, P.; and Sander, C. 1995. Transmembrane helices predicted at 95% accuracy. *Protein Sci*. 4(3):521–533.

Rost, B.; Fariselli, P.; and Casadio, R. 1996. Topology prediction for helical transmembrane proteins at 86% accuracy. *Protein Sci*. 5(8):1704–1718.

Schwarz, R., and Chow, Y.-L. 1990. The N-best algorithm: An efficient and exact procedure for finding the N most likely hypotheses. In *Proceedings of ICASSP'90*, 81–84.

Sipos, L., and von Heijne, G. 1993. Predicting the topology of eukaryotic membrane proteins. *Eur. J. Biochem.* 213(3):1333–1340.

Traxler, B.; Boyd, D.; and Beckwith, J. 1993. The topological analysis of integral cytoplasmic membrane proteins. *J Membr Biol* 132:1–11.

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

von Heijne, G. 1992. Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. *Journal of Molecular Biology* 225(2):487–494.

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

Wallin, E., and von Heijne, G. 1998. Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci.* in press.

Wallin, E.; Tsukihara, T.; Yoshikawa, S.; von Heijne, G.; and Elofsson, A. 1997. Architecture of helix bundle membrane proteins: an analysis of cytochrome c oxidase from bovine mitochondria. *Protein Sci*. 6(4):808–815.