

Methods for Transmembrane Protein Topology and Alpha Helix Prediction

Kristen Carnohan
BIOC 218: Computational Molecular Biology
Final Project
December 9, 2012

Introduction

Alpha-helical transmembrane proteins are essential to many biological processes, such as transport, signaling, intracellular communication, cell recognition, and adhesion¹. These proteins also comprise the majority of drug targets. However, because experimentally determining the structures of these molecules is often slow and difficult, relatively very limited experimental structural data for membrane proteins is available in protein data banks. Therefore the ability to instead accurately predict the topology of transmembrane proteins is inherently useful. Any paper on the prediction of transmembrane protein topology will begin by agreeing that the problem of characterizing protein structures is a critically important endeavor in computational biology, and that it is one of the more difficult problems to solve¹⁻⁹. Accordingly, there are many available programs that conduct such predictions utilizing several different approaches. The strategies of these methods can generally be categorized into two types: those that use a residue-based analysis to determine the likelihood of each amino acid to appear in each protein region, and those whose goal is to match an overall model of a protein to the given amino acid sequence².

This paper explores five different methods for the prediction of transmembrane protein topology and alpha helices, in the chronological order that they were developed. These five methods are: TopPred (1992), MEMSAT (1994), PHDhtm_ref (1996), HMMTOP (1998), and TMHMM (1998). Next, the accuracy of these methods is examined, and the question of which is the superior of the five is discussed.

Algorithms

TopPred: Hydrophobicity and the Positive-Inside Rule

Gunnar von Heijne pioneered a simple method of transmembrane topology prediction in 1992³. This method is the basis for the program TopPred (sometimes referred to as TOPPRED or TOP-PRED), short for “topology prediction.” At the time of the TopPred method’s development, the standard method for transmembrane topology prediction was a simple hydrophobicity analysis. von Heijne’s method utilizes and builds

upon this strategy, adding a step of charge-bias analysis to rank all possible structures using the positive-inside rule. His method is highly successful at predicting the topology of bacterial inner membrane proteins.

The first step to reaching a prediction using the TopPred method is to compose a list of all possible transmembrane segments in the given protein using hydrophobicity analysis³. A hydrophobicity profile is formulated using the GES-scale (Engelman *et al.*, 1986). von Heijne used a trapezoidal sliding window (a parameter for calculating the hydrophobicity profile) in favor of the commonly-used triangular and rectangular windows, since the trapezoid combined the other shapes' respective strengths of noise reduction and realism.

Candidate transmembrane segments are extracted from the hydrophobicity profile by identifying the highest peaks in hydrophobicity above a certain (fairly lenient) cutoff (Heijne used a cutoff of 0.5 for this step)³. Of the list of potential transmembrane segments, those with a hydrophobicity of 1.0 or greater can be deemed "certain" transmembrane segments, while those with hydrophobicity between 0.5 and 1.0 remain possible, but not definite, candidates. It is this set of uncertainly classified segments that the charge-bias screening step (to follow) works to define. The cutoff numbers used in the method were derived from analysis of transmembrane proteins with experimentally verified topologies.

Next, a list of all possible topologies of the protein is automatically generated using a computer program³. At this point, all the possible topologies must include every definite transmembrane segment, but may either include or exclude each of the tentative segments.

The next stage of the method is based on the positive-inside rule. The positive-inside rule can be summarized as follows: transmembrane proteins tend to have a much higher concentration of positively charged amino acids on the cytoplasmic side of the membrane than on the periplasmic side (the periplasm is the space between the inner and outer membranes of Gram-negative bacteria)³. To discern whether a protein conforms to the positive-inside rule, we look to the charge-bias. The charge-bias of a protein is a measure of the difference in charge between its cytoplasmic and periplasmic segments. A more positive charge-bias indicates greater compliance with the positive-inside rule, which is most often a harbinger of a better prediction. [Figure 1](#) illustrates supporting evidence that the positive-inside rule holds for bacterial inner membrane proteins.

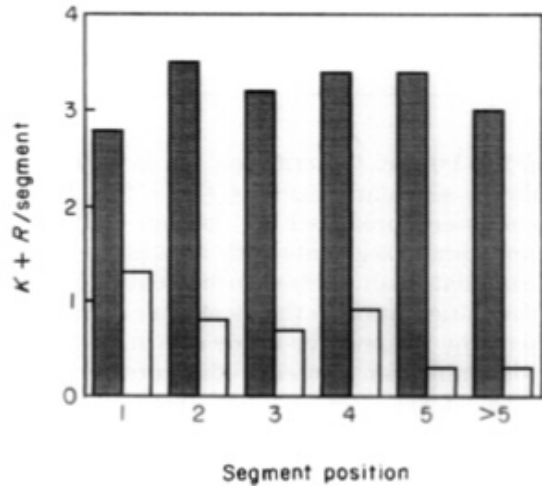


Figure 3. Mean number of Lys+Arg in cytoplasmic (stippled bars) and periplasmic (open bars) polar segments as a function of the position of the segment (1st cytoplasmic segment, 2nd cytoplasmic segment, etc., counting from the N terminus) in a sample of 24 bacterial inner membrane proteins with known topology.

[Figure :](#) Chart supporting the positive-inside rule, from von Heijne's 1992 paper introducing the prediction method used by TopPred³.

With the positive-inside rule in mind, the charge-bias is calculated for each of the contending topologies, which are then ranked in order of decreasing charge-bias³. Each structure is oriented such that its more highly charged side faces the cytoplasm. The top-ranked structure is chosen as the method's final prediction of the protein's true topology. This is where the positive-inside rule comes in; a structure with an erroneous number of transmembrane segments has at least one polar domain on the incorrect side of the membrane. This would likely cause the false structure to have a lesser charge-bias than the true structure. In this way the positive-inside rule can protect the TopPred method from predicting an incorrect structure in most cases. [Figure 2](#) depicts an example of how the positive-inside rule can help distinguish between accurate and faulty topology choices.

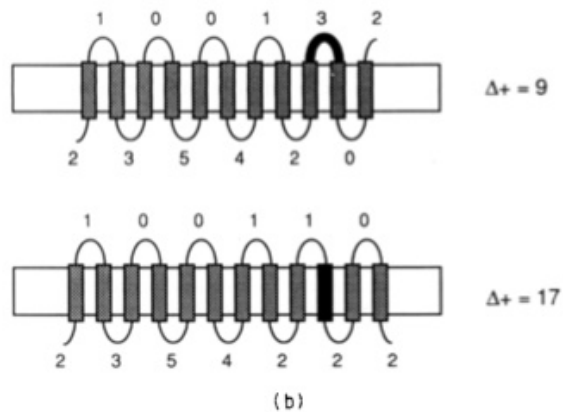
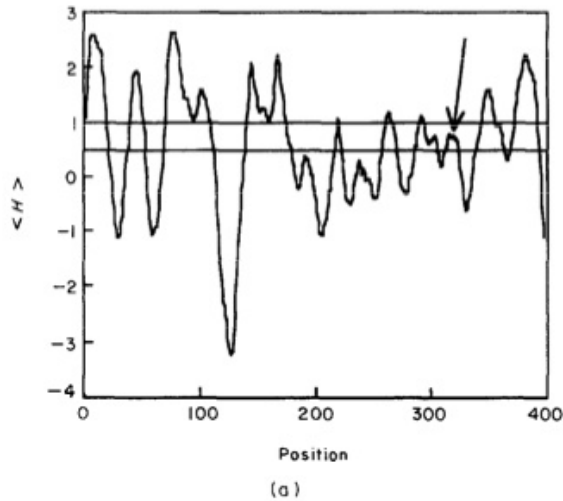


Figure 4. (a) Hydrophobicity plot for the SecY protein. The upper and lower cutoffs are marked. A tentative transmembrane segment with a mean hydrophobicity falling between the 2 cutoffs is marked by an arrow. (b) Two possible topologies for the SecY protein based on the hydrophobicity plot. The putative transmembrane segment is shown in black. The number of Arg+Lys residues is shown next to each polar segment. Note that the correct alternative (bottom, including the putative transmembrane segment) has a much higher charge-bias than the incorrect one.

[Figure](#) : An example of how using charge-biases can help determine the best prediction between two candidate structures³.

MEMSAT: Dynamic Programming

The server MEMSAT employs a transmembrane topology modeling method, developed in 1994 by Jones, Taylor, and Thornton. This approach uses a set of log-likelihood tables compiled from data on well-characterized transmembrane proteins, and includes a dynamic programming algorithm that implements expectation maximization to

predict topology models⁴. Expectation maximization seeks to find a model which best explains the observed data.

In order to conduct expectation maximization, first a statistical model is defined which includes parameters for number of transmembrane segments, topology (delineates whether the N-terminus is inside the cell or outside), length, and location of each segment within the total protein sequence⁴. In this model there are five structural states (Figure 3): inside loop (L_i), outside loop (L_o), inside helix end (H_i), middle helix (H_m), and outside helix end (H_o). The length in residues of helix ends was arbitrarily set to 4.

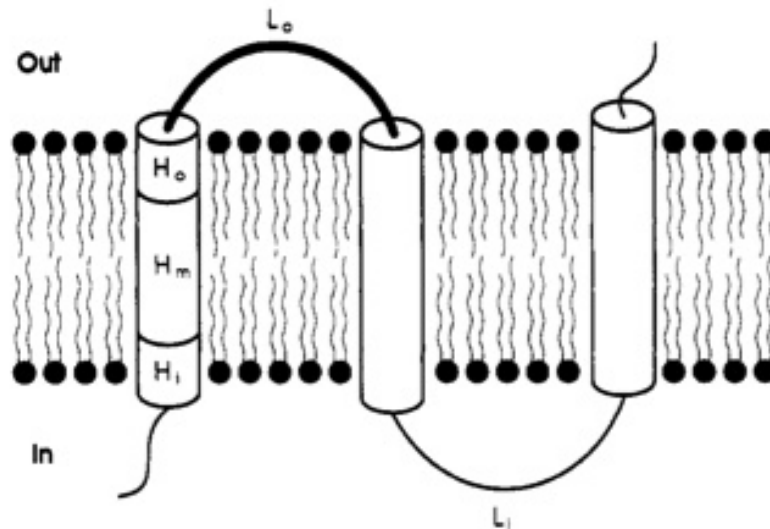


FIGURE 1: Structural states defined for a typical helical transmembrane protein.

[Figure 3: structural states for the MEMSAT method⁴.](#)

It is well-known that cytoplasmic, transmembrane, and extracellular protein segments each have differing observed biases toward particular amino acids, and this fact can be used to derive a topology prediction from an amino acid sequence⁴. Quantifying these biases using a log likelihood calculation can leverage them for use in modeling protein topologies. For each of the five structural states, the log likelihood ratios of each of the 20 amino acids were formulated using the equation

$$s_i = \ln(q_i/p_i)$$

where s_i is the log likelihood of amino acid i in a particular state; p_i is the frequency of amino acid i out of all the amino acids in all the proteins of the data set; and q_i is the relative frequency of amino acid i in the given structural state. A score near to zero indicates that the frequency of the amino acid in the given state is the same as the expected frequency by chance alone. A positive ratio signifies higher than random frequency, while a negative ratio implies a frequency lower than that obtained by chance.

The log likelihood scores are used as parameters to calculate a score indicating whether a given protein sequence is compatible with a particular topology model⁴. For a single residue, the score is dependent on the identity of the residue and in which of the five structural states it resides. In order to define the highest-scoring set of transmembrane helix positions and lengths, MEMSAT uses a recursive dynamic

programming algorithm that is almost identical to the Needleman-Wunsch algorithms used for pairwise sequence alignment.

As in the Needleman-Wunsch algorithms, a score matrix is formulated (Figure 4) to predict the best topology⁴. For any given protein, two scoring matrices must be defined in order to enumerate every possible topology model, since the N-terminus of the protein can lie on either the cytoplasmic or extracellular side of the cell membrane. The dynamic programming algorithm calculates the highest-scoring path possible from the matrices, and the corresponding topology is selected as the final prediction (Figure 5).

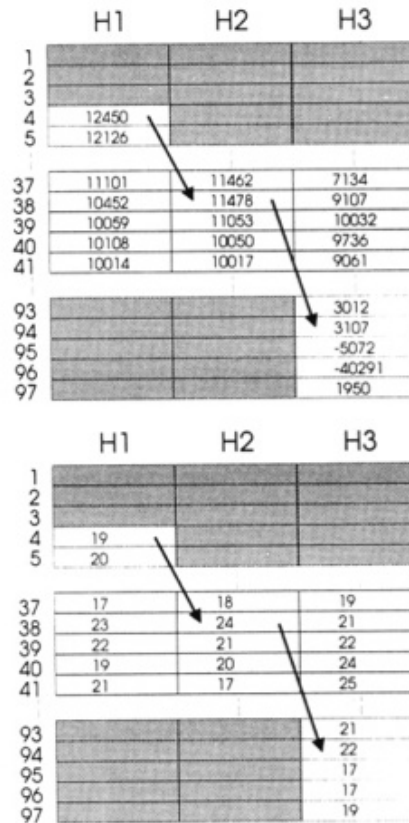


FIGURE 4: A hypothetical score matrix for three transmembrane helices. The upper matrix holds the highest achievable path score for each cell, and the lower matrix stores the helix length which permits this score.

Figure 4: The scoring matrix method employed by the MEMSAT prediction program⁴.

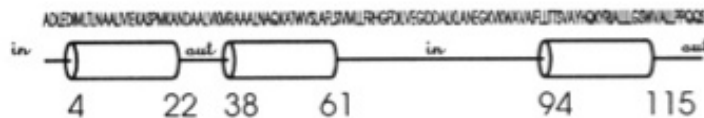


FIGURE 5: Predicted structure and topology relating to the optimal path shown in Figure 4.

Figure 5: The translation of the highest matrix score into a state model of a predicted transmembrane protein⁴.

PHDhtm_ref: Neural Networks

Rost developed a method for identifying transmembrane helices in 1996 that relies on the use of neural networks, named PHDhtm_ref⁵. The general idea is to feed a multiple sequence alignment to a system of layered neural networks.

The first step in the PHDhtm_ref method is to generate the multiple sequence alignment, which should possess a high level of accuracy and contain a wide range of homologous sequences for the method to perform optimally⁵. From there, the sequence alignment and subsequent inputs are processed through several different levels. The structural state of a transmembrane protein segment can be a helix, a strand, or a loop, and the segment can either be transmembrane or non-transmembrane.

The primary level is a neural network whose input consists of a local sequence window of 13 adjoining residues and the global sequence⁵. This sequence-to-structure network outputs the 1D structural state of the central residue in the input window.

The output of the first level serves as the input to the secondary level, which is another neural network⁵. This one is a structure-to-structure network with the same output units as in the first level. Because the sample proteins used to train the networks are selected randomly, the examples from one time step to the next are normally not adjacent to one another in sequence, which inhibits the first level network's ability to learn length distributions. Thus the second network is necessary to incorporate correlations between adjoining residues, which make predicted segments and helices of realistic length possible. A visual representation of the PHDhtm_ref method up to this point in the process can be viewed in [Figure 6](#).

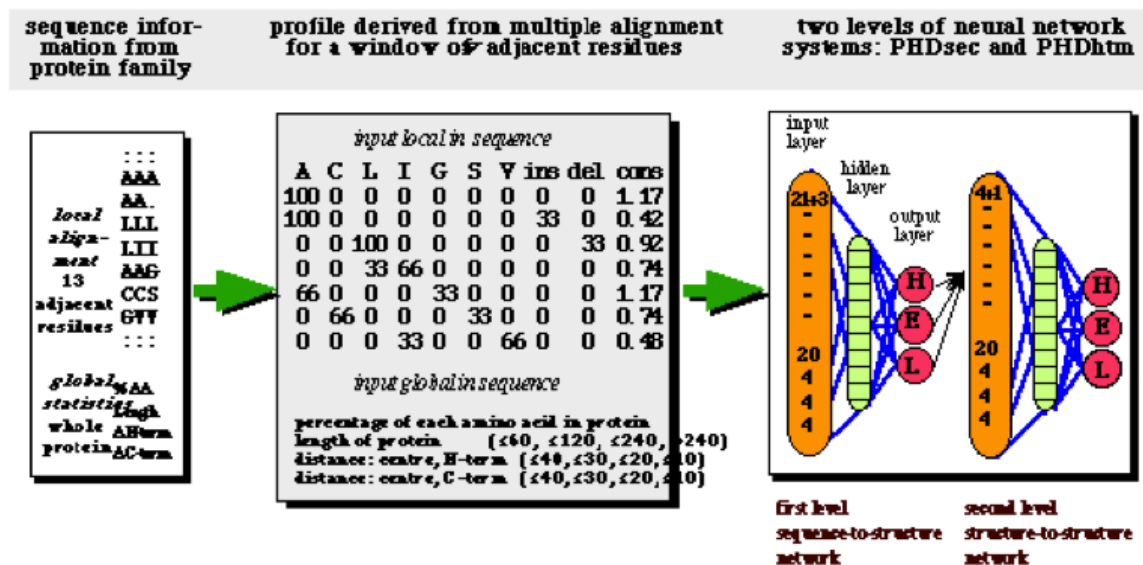


Figure : The workflow of the PHDhtm_ref method for transmembrane topology prediction⁵.

The third level is simply the arithmetic average over independently trained networks. Some networks are trained on protein segment states in their natural proportions, but this causes bias towards some states with high representation (such as

loops) over others that are less frequent (like transmembrane helices)⁵. This “unbalanced training” is complemented by the “balanced training” of other neural networks, wherein the network is presented with a set of examples possessing even proportions of all the states. Because one training method tends to overpredict a given state while the other underpredicts, the average is a compromise designed to calculate the correct number of segments for each state. This “jury decision” is typically a simple arithmetic mean over four networks: level 1 unbalanced, level 1 balanced, level 2 unbalanced and level 2 balanced.

The fourth and final level is simply a filter to weed out impossibilities in predicted topologies. It either splits or shortens transmembrane helices that are too long, and lengthens or deletes those that are too short⁵. Once the predicted model has passed through the filter, it is outputted as the final prediction.

HMMTOP –A Hidden Markov Model

In 1998, Tusnady and Simon developed a hidden Markov model prediction method. This method, unnamed at the start, is the method used by the HMMTOP server, and will therefore be referred to as “HMMTOP” or “the HMMTOP method” from this point forward. Its creators described the method as being “based on the hypothesis that the localizations of the transmembrane segments and the topology are determined by the difference in the amino acid distributions in various structural parts of these proteins rather than by specific amino acid compositions of these parts.”⁶ Thus, the success of the method provided both a new and fairly accurate prediction tool for transmembrane protein topology, and a support for its creators’ hypothesis about the theory of transmembrane protein topology itself.

The HMMTOP method uses a hidden Markov model to find the most probable of all the possible topologies of a protein, which is a prediction for and hopefully a match with the experimentally determined topology. HMMTOP’s HMM is comprised of five structural states: inside loop, inside helix tail, membrane helix, outside helix tail, and outside loop⁶. A “loop” is defined as a sequence of amino acids outside the membrane. A “tail” is a section of a membrane helix that protrudes from the membrane into the cytoplasm or extracellular matrix of the cell. A membrane helix is always sandwiched between two tails in the model. A tail can be immediately followed by either a loop or another tail, forming a “short loop” comprised of two tails. [Figure 7](#) provides a visualization of these structural states.

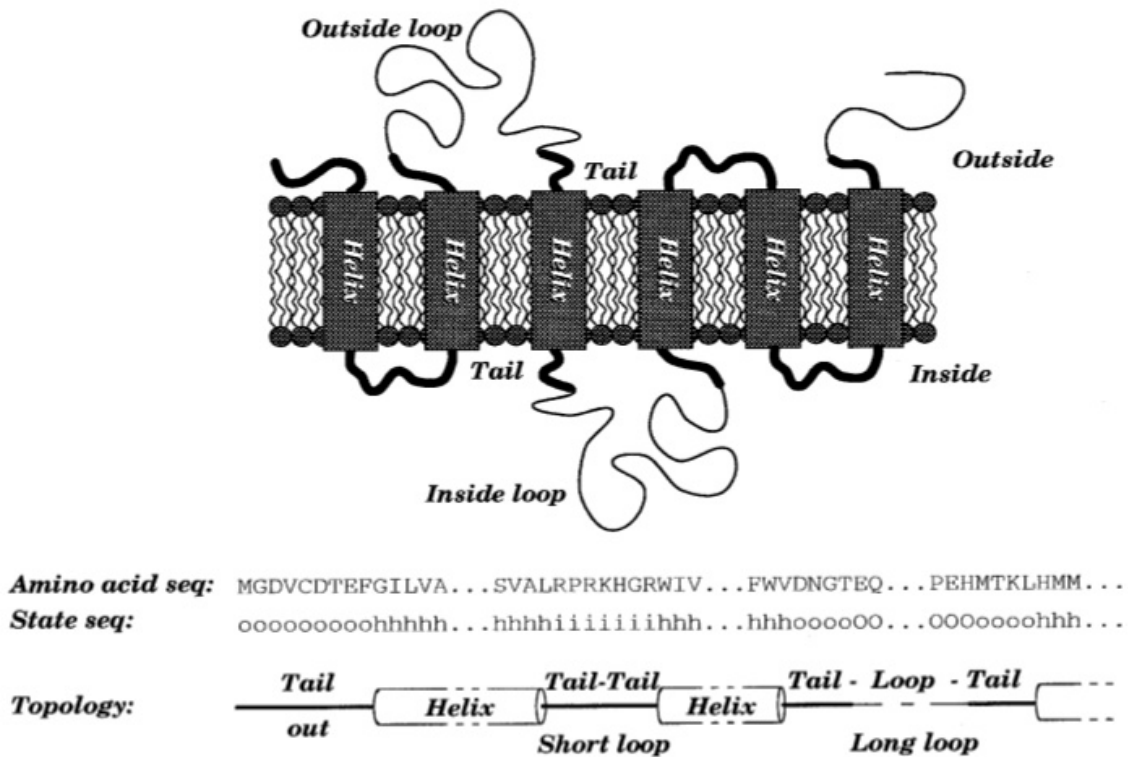


Figure 2. Structural states defined for a typical helical transmembrane protein. The five states are: inside loop (l), inside tail (i), membrane helix (h), outside tail (o) and outside loop (O). Tails (thick lines) are thought to interact with the inside or outside parts of the membrane, while loops (thin lines) do not. Two tails between helices can form a short loop, but longer loops are formed by tail-loop-tail sequences.

Figure : Illustration and caption explaining the biological basis for the HMMTOP model's five states⁷.

HMMTOP defines two state types based on the observation that short loops (5-30 residues) and long loops (more than 30 residues) generally have different length distributions⁶. These two state types are labeled non-fixed length (NFL) and fixed-length (FL). The difference between the two state types is the number of possible transitions from a NFL state versus a FL state; a NFL can be succeeded either by the same state, which adds length to that state, or by a transition to the next state. This reflects the observed geometric length distribution of long loops, which can be of arbitrary length. On the other hand, a FL state is divided into MAXL states, which constrains its length between MINL and MAXL. From the first MINL substates, the only possible transition leads to the next substate, while between MINL and MAXL, a second transition is available to leave the current state entirely and move to the next one. The transitions between the substates have varying probabilities. Loops are categorized as NFL states, while tails and helices are FL.

The creators of the HMMTOP method designed the progression from one state to the next to follow the natural state progressions of transmembrane proteins. [Figure 8](#) illustrates the state architecture of the HMM; notice that the possible transitions between states do in fact reflect the state transitions in a transmembrane protein like the example pictured in [Figure 7](#).

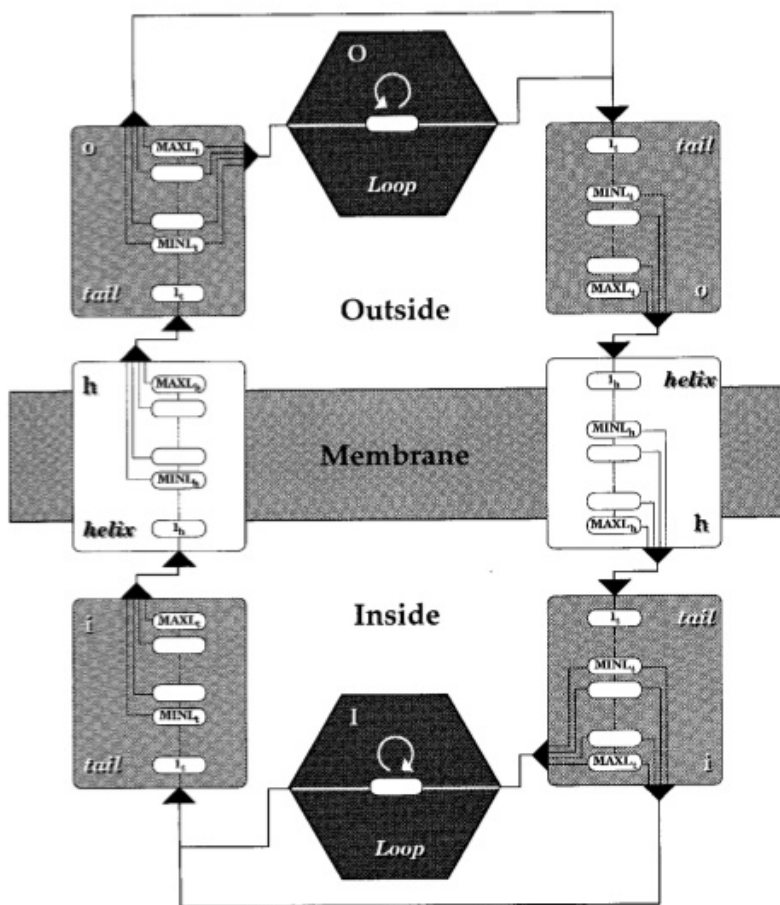


Figure 3. Architecture of HMM used for topology prediction. States with the same transition matrices are colored in the same way: white, helix states; light gray, tail states; dark gray, loop states. Rectangular areas FL type states; hexagonal ones, NFL type states. The observation-symbol probabilities used by states are marked in each state. The structure of substates in the case of the FL type is drawn within states. Lines and arrows show the possible transition between states or sub-states.

Figure 3: Figure representing the basic HMM architecture of the HMMTOP model, along with an explanation of color-coding and shapes used in the figure.

Before the model can be used to make a prediction, preliminary estimates of HMM parameters must be set. These can be derived either randomly or based on predetermined values, and are then optimized for the given amino acid sequence or its homologs⁶. Then the best-fitting state sequence is calculated given the HMM and its parameters. The parameters the creators used to train the model were based on transmembrane proteins with experimentally validated topologies, and could be derived either from a single sequence or from multiple sequences; training with multiple sequences tends to increase the model's accuracy during testing.

TMHMM –Another Hidden Markov Model:

Sonhammer, Von Heijne, and Krogh developed TMHMM in 1998. This method is based on a different hidden Markov model with seven states rather than five, and can distinguish between soluble and membrane proteins⁷.

Figure 9 represents the architecture of TMHMM's model. In (A) each box represents a section of the structure that models a particular region of a transmembrane

protein (helix caps, center of a helix, areas near the membrane, and globular domains). Each box with the same name in the diagram shares the same parameters⁸. There are two different models for non-cytoplasmic loops, one for short and the other for longer loops. This is because short and long loops are the two possible membrane entry mechanisms, and the two types of loops have distinct properties from one another.

In turn, each region model contains a set of HMM states, which model the lengths of their respective regions. (B) and (C) of Figure 9 illustrate state structure diagrams for helix core, globular, loop, and cap regions. The arrows in the figure represent transitions from one state to another that are acceptable according to the grammatical structure of helical transmembrane proteins⁷. Each state in a submodel is tied to specific amino acid probabilities; these parameters are the same for all the states in one region, but the parameters of different boxes are varying⁸.

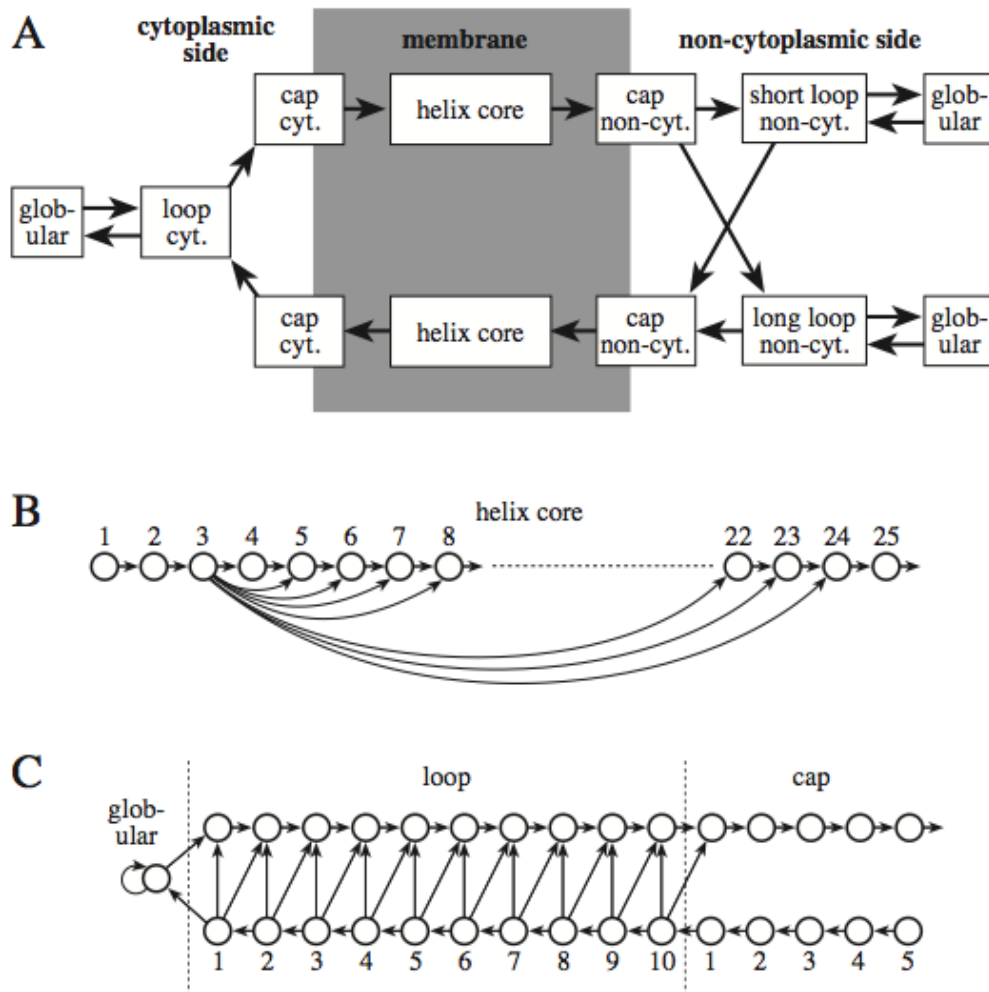


Figure 9: (A) The overall structure of the TMHMM model⁸. Each box represents a specific region of a transmembrane protein. (B) Submodel of states for the helix core region. (C) Submodel of states for globular, loop, and cap regions

The HMM is trained on a set of transmembrane proteins with known topology, including verified locations of transmembrane alpha-helices⁷. The total number of HMM

parameters tracked by the model is 216, which is significantly less than that of neural network models, which usually incorporate parameters in the tens of thousands⁸.

TMHMM predicts transmembrane helices by calculating the most likely overall topology based on the HMM. However, since there is often some uncertainty about the location of the helices –whether they are embedded in the membrane, enclosed in the cytoplasm, or in the extracellular matrix –all three of these probabilities can be used to display alternative topologies and their relative probabilities⁷ (see [Figure 10](#)).

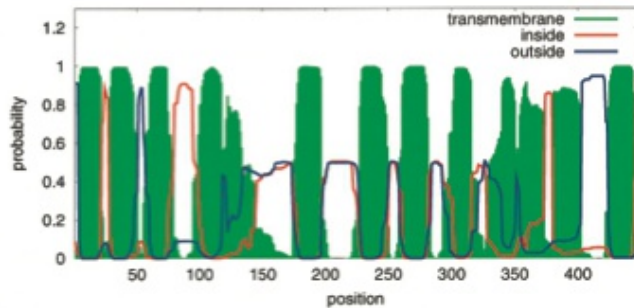


Figure 2. Posterior probabilities for a single sequence. The posterior probability for transmembrane helix, inside, or outside displayed for the gluconate permease 3 from *E. coli* (SWISS-PROT entry Gntp_ECOLI), for which the structure is unknown. Some parts of the protein are relatively certain, whereas other parts are less certain. It is unclear, for instance whether there are one or two transmem-

brane segments between amino acid 100 and 150, and between 325 and 375. This uncertainty is also reflected in a total uncertainty in which side the loops are (inside or outside) between 150 and 325. For this protein the single most probable topology turns out to have two helices in both of these regions giving 13 transmembrane helices in total, and this prediction turns out to be essentially identical to the annotation in SWISS-PROT. However, the posterior probability plot shows that the topology with only one helix in these regions (11 in total) is a quite likely alternative, whereas a topology with 12 or 14 transmembrane helices is not so likely because it would fit badly with the posterior probabilities of inside/outside in the two ends of the protein. In Klemm *et al.* (1996) 14 transmembrane helices are predicted for this protein; three helices are predicted in the region between 100 and 150.

Figure 10: Figure and caption illustrating the concept of posterior probabilities, wherein all three probabilities of alpha-helix location are considered when predicting the overall protein topology⁷.

One benefit conferred by the use of a hidden Markov model is TMHMM's ability to model the length of a transmembrane helix⁷. The distribution of alpha-helix core lengths in transmembrane proteins with known structure ranges from 5 to 25 residues long. The helix length possibilities in TMHMM's model precisely mirror this biological rule, as is evident from the helix core model in [Figure 1 \(B\)](#). All helix core lengths produced by the model are between 5 and 25, and whole helices are between 15 and 35 residues in length when the two caps are added⁸.

The model's accuracy stems from the fact that it is a close mapping of biological realities. Not only does TMHMM take into account helix length, but it can also incorporate hydrophobicity, charge bias, and grammatical limitations, all in a single integrated model⁷.

Accuracy of Methods

Each developer or team of developers, in their introductory papers, presents a test of their methods and reports on estimated accuracy. These reports are presented here, followed by a discussion of *true* accuracy measured by outside parties.

von Heijne tested his TopPred method on bacterial inner membrane proteins whose sequences and topologies were experimentally well defined³. It correctly predicted the overall topologies of 23 out of 24 proteins (96% accuracy), and identified all 135 transmembrane segments from the sample, plus one overprediction.

The MEMSAT method successfully predicted 64 out of 83 entire topologies tested (yielding an accuracy of about 77%), including 34 correct of 37 complex multispanning proteins⁴.

With the PHDhtm_ref method, Rost was able to correctly predict 365 transmembrane helices out of 380 helices in 69 test proteins⁵. This corresponds to an approximate accuracy of 96%. Tested the opposite way for overprediction, given 278 globular non-transmembrane proteins, PHDhtm_ref only predicted 14 incorrect transmembrane helices.

HMMTOP achieved about 96% average accuracy in predicting transmembrane helices in data sets, and was able to predict the overall topology correctly in the same data sets with an average accuracy of 85%⁶.

The creators of the TMHMM method indicate that, in cross-validated tests on sets of 83 and 160 proteins with known topology, their method was successful in predicting the entire topology of a protein 85% of the time for both data sets⁸ (Figure 11).

Table 1. Results on various data sets using HMM for topology prediction

Data set	No. of transmembrane helices						No. of correct proteins		
	N_{obs}	N_{prd}	N_{cor}	Q_P (%)	Q_2	N_{TOT}	N_{TM}	N_{TT}	Q_T (%)
83TMP	346	353	344	98.4	94.9	83	74	72	87
48TMP	194	197	194	99.2	94.6	47	45	43	91
prokTMP	262	264	259	98.5	89.8	44	38	32	73
Total	698	709	694	98.7	94.2	158	143	135	85

N_{obs} , N_{prd} and N_{cor} are the number of observed, predicted and correctly predicted transmembrane helices, respectively; $Q_P = 100 \cdot \sqrt{(N_{cor}/N_{obs}) \cdot (N_{cor}/N_{prd})} \cdot N_{TOT}$, N_{TM} and N_{TT} are the number of proteins in the data sets, the number of proteins for which all transmembrane segments were predicted correctly, and the number of proteins for which both the topology and the transmembrane segments were predicted correctly, respectively. Q_2 is the per residue accuracy.

[Figure](#) : reported test results for TMHMM⁸.

The developers of TMHMM, when testing their model, also tested other methods using the same data sets⁸, and a resulting table of accuracy is provided in [Figure 12](#). Four out of the five approaches studied in this paper are included; only HMMTOP was left out. Not surprisingly, the three other methods tested against TMHMM all performed more poorly than the multi-sequence version of TMHMM in predicting overall topologies, and

all did worse in these tests than in those performed by their own developers (60% vs. 96% for TopPred, 65% vs. 77% for MEMSAT, and 81% vs. 96% for PHDhtm_ref).

Table 2. Prediction accuracy of various algorithms on various data sets

Data set	Method	No. of transmembrane helices				Q_P (%)	N_{TOT}	No. of correct proteins		Q_T (%)
		N_{obs}	N_{pred}	N_{cor}	N_{TM}			N_{TT}		
83TMP	TOPPRED	346	381	336	92.5	83	54	54	65	
	MEMSAT		351	336	96.4		69	65	78	
	HMM ₁		358	342	97.2		68	66	80	
	PHDhtm_ref		351	342	98.1		75	73	88	
	HMM _{multi}		353	344	98.4		74	72	87	
48TMP	MEMSAT	194	174	165	89.8	47	26	23	49	
	TOPPRED		200	193	98.0		40	25	53	
	HMM ₁		198	192	98.0		40	39	83	
	HMM _{multi}		197	194	99.2		45	43	91	
	PHDhtm_ref		192	192	99.5		45	42	89	
prokTMP	PHDhtm_ref	262	259	254	97.5	44	31	28	64	
	MEMSAT		255	250	96.7		33	29	66	
	TOPPRED		264	255	97.0		32	30	68	
	HMM ₁		264	258	98.1		36	30	68	
	HMM _{multi}		264	259	98.5		38	32	73	
Total	TOPPRED	698	740	681	94.8	158	112	95	60	
	MEMSAT		673	647	94.4		114	103	65	
	HMM ₁		714	689	97.6		131	124	78	
	PHDhtm_ref		699	685	98.1		136	128	81	
	HMM _{multi}		709	694	98.7		143	135	85	

References for methods are as follows: TOPPRED (von Heijne, 1992), MEMSAT (Jones *et al.*, 1994), PHDhtm_ref (Rost *et al.*, 1996), HMM₁ hidden Markov model used on single sequence information in this article, HMM_{multi} hidden Markov model used on multiple sequence information in this article. The meanings of the columns are the same as in Table 1.

Figure : Accuracy of TopPred, MEMSAT, PHDhtm_ref, and TMHMM, as measured by the developers of TMHMM⁸.

Which of these methods is really the most precise, and why are there discrepancies in reported accuracy? Rost contends that many creators of prediction methods tend to overstate their models' accuracies by using questionable or biased testing procedures, such as using only tens of proteins in sample sets used to estimate accuracy⁵, and indeed most of the developers of methods studied here are guilty of using such small sample sizes.

The inconsistency in general stems from the fact that there is no standard data set or procedure for testing the prediction methods. Viklund and Elofsson agree, stating, "Differences between the evaluations are due to what is being measured (per residue accuracy, per protein accuracy, etc.) and perhaps more important, the composition of the data set used in the comparison, which may be more or less similar to the data set for which a particular method has been optimized."²

Even when tested by an outside party with standardized data sets in order to more robustly determine the superior method, none of the prediction programs surfaced as the clear winner. Chen *et al.* found that "some methods are better; none are clearly best... no method(s) performed consistently better than all others by more than one standard error."⁹ The fact that no model significantly outperformed all the others is a clue that the "perfect" prediction method is still on the horizon.

Although no method can be definitively proclaimed as the best for all cases, some types of models do seem to be on the right track. In a 2004 study, Viklund and Elofsson found that HMM methods (such as TMHMM and HMMTOP) were generally more accurate than residue profiling strategies utilized by earlier programs². When these HMMs were trained on multiple sequences, and when evolutionary information was additionally accounted for, these methods performed even better.

References

- (1) T. Nugent and D. Jones (2009). "Transmembrane protein topology prediction using support vector machines." BMC Bioinformatics 10: 159.
- (2) H. Viklund and A. Elofsson (2004). "Best - helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information." Protein Science 13, 7: 1908-1917.
- (3) G. von Heijne (1992). "Membrane Protein Structure Prediction: Hydrophobicity Analysis and the Positive-inside Rule." Journal of Molecular Biology 225: 487-494.
- (4) D. T. Jones, W.R. Taylor, and J.M. Thornton (1994). "A Model Recognition Approach to the Prediction of All-Helical Membrane Protein Structure and Topology." Biochemistry 33: 3038-3049.
- (5) B. Rost (1996). "PHD: predicting 1D protein structure by profile based neural networks." Meth. in Enzym. 266: 525-539.

- (6) G.E. Tusnady and I. Simon (1998). "Principles Governing Amino Acid Composition of Integral Membrane Proteins: Application to Topology Prediction." Journal of Molecular Biology 283: 489-506.
- (7) A. Krogh, B. Larsson, G. von Heijne, and E. L. L. Sonnhammer (2001). "Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes." Journal of Molecular Biology 305: 567-580.
- (8) E.L.L. Sonnhammer, G. von Heijne, and A.Krogh (1998). "A hidden Markov model for predicting transmembrane helices in protein sequences." Proc Int Conf Intell Syst Mol Biol 6: 175-182.
- (9) C.P. Chen, A. Kernytsky, and B. Rost (2002). "Transmembrane helix predictions revisited." Protein Science 11, 12: 2774-2791.