The Use of Internal and External Functional Domains to Improve Transmembrane Protein Topology Prediction

by

Emily Wei Xu

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirement for the degree of

Master of Mathematics

in

Computer Science

Waterloo, Ontario, Canada, 2004

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Membrane proteins are involved in vital cellular functions and have important implications in disease processes, drug design and therapy. However, it is difficult to obtain diffraction quality crystals to study transmembrane protein structure. Transmembrane protein topology prediction tools try to fill in the gap between abundant number of transmembrane proteins and scarce number of known membrane protein structures (3D structure and biochemically characterized topology). However, at present, the prediction accuracy is still far from perfect. TMHMM is the current state-of-the-art method for membrane protein topology prediction. In order to improve the prediction accuracy of TMHMM, based upon the method of GenomeScan, the author implemented AHMM (augmented HMM) by incorporating functional domain information externally to TMHMM. Results show that AHMM is better than TMHMM on both helix and sidedness prediction. This improvement is verified by both statistical tests as well as sensitivity and specificity studies. It is expected that when more and more functional domain predictors are available, the prediction accuracy will be further improved.

Acknowledgements

The research is supervised by Dr. Paul Kearney along with help from Dr. Daniel G. Brown in an early course project. My readers are Dr. Daniel G. Brown and Dr. Ming Li.

I would also like to thank my colleagues Ms. Brona Brejova, Mr. John Tsang and Mr. Mike Hu from the University of Waterloo, Dr. Peter Ehlers and Dr. Andrei Turinsky from the University of Calgary for their helpful discussions and Dr. Michel Dominguez from Caprion for his opinion on functional domain sidedness.

This thesis is dedicated to God for His love, mercy and grace.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

About 20% to 25% of proteins are membrane proteins [1, 2, 3]. Of particular interest are cell surface integral membrane proteins, since they have significant implications in disease processes, drug design and therapy. However, it is very difficult to crystallize membrane proteins to study their structures. Thus, reliable prediction of the topology of transmembrane (TM) proteins from amino acid sequence is an important tool in protein research. Topology refers mainly to the location, number and orientation of the membrane spanning segments. TMHMM (TransMembrane Hidden Markov Model) is the best prediction program so far for membrane protein topology [1]. Unless it is specified otherwise, TMHMM refers to both TMHMM 1.0 and TMHMM 2.0. However, it has less than 52% accuracy on the prediction of TM proteins collected by Moller *et al*. [1, 4]. The improvement for the sidedness (orientation) of TM proteins remains a priority since the prediction accuracy for sidedness is even lower than the prediction accuracy for helix location. Furthermore, accurate sidedness prediction enables cell surface epitopes to be predicted for immunotherapies.

The following chapters approach membrane protein topology prediction from both biological and computational standpoints. Chapter 1 gives a brief introduction to current research on TM protein topology prediction and biological background on TM proteins. Chapter 2 focuses on current prediction methods and potential improvement approaches. Chapter 3 presents experiments and results and Chapter 4 brings forth discussions and future work.

This chapter not only serves to provide the framework for biological background, but also helps to introduce some other key aspects or problems that membrane protein topology prediction programs must deal with. For example, program SignalP [13] deals with the distinction between signal peptide and signal anchor. In addition, a good prediction tool has to be able to distinguish between α-helical and β-barrel membrane proteins. Needless to say, transmembrane protein assembly is a very complicated biological process. The full mechanism has not been fully elucidated. However, transmembrane protein assembly is the starting point for membrane protein topology prediction because during assembly, the signals embedded within the protein sequences are decoded. In this process, the transmembrane protein is directed to the correct location of the cell and helps it assume its proper topology.

1.1 Cell membrane and transmembrane proteins

Every cell is bounded by a cell membrane (Figure 1 shows a liver cell [5]). For brevity, in this thesis, the cell refers to a eukaryotic cell unless otherwise stated.

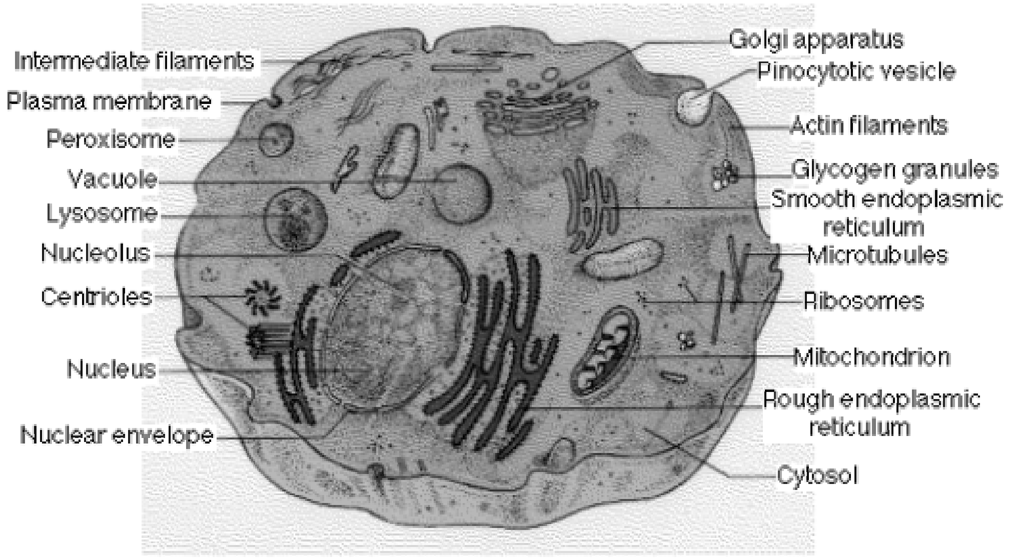

Figure 1: A generic representation of a typical eukaryotic cell (liver cell) bounded by a cell membrane with organelles inside.

The cell membrane is also called the plasma membrane (PM). The plasma membrane is composed of a lipid bilayer (two layers of lipids) and associated proteins, which include integral membrane proteins and peripheral membrane proteins (Figure 2). Integral membrane proteins are often referred to as transmembrane proteins. We are especially interested in integral membrane proteins, because they are involved in vital cellular functions such as cell-cell communication, recognition, adhesion, membrane fusion, and transportation. They include transport proteins, receptors, and enzymes, for example.

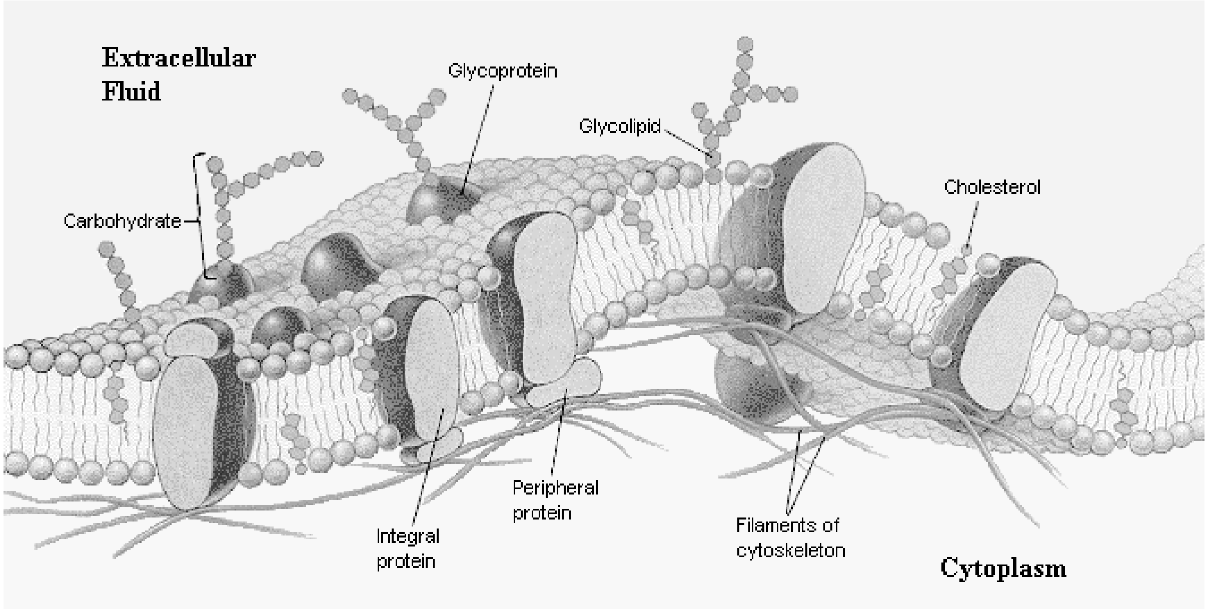

Figure 2: Graphical illustration of integral and peripheral membrane proteins in eukaryotic cell membranes (taken from [42]).

There are two known classes of integral membrane proteins: those with α-helical structure and those with β-barrel structure. Since at present there are only 12 β-barrel sequences with known structure, α-helical structure is our modeling focus. From the observation of the 3D α-helical structures of bacteriorhodopsin (determined by electron diffraction, Figure 3 [6]) and the photosynthetic reaction center (determined by X-ray crystallography), researchers conclude that transmembrane segments are 17-25 residues long apolar helices. Prediction programs for integral membrane proteins usually assume that the helices completely traverse the membrane and are perpendicular to its surface [7].

β-barrel TM proteins comprise of even numbers of β-strands [8]. Figure 4 shows the 3D β-barrel structure of TolC outer membrane protein of *E. coli* [6]. Although β-barrel membrane spanning regions generally are shorter and much less hydrophobic than those in α-helical membrane proteins, they could still be a source of false positives, and be predicted as α-helical membrane proteins [2].

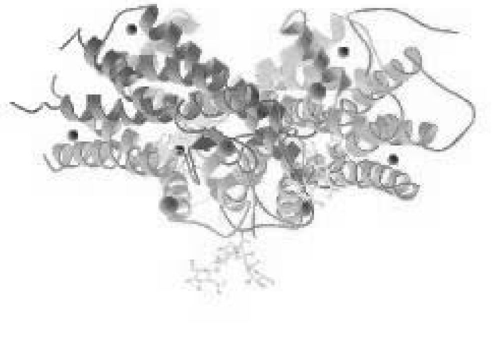

Figure 3: α-helical structure of bacteriorhodopsin.

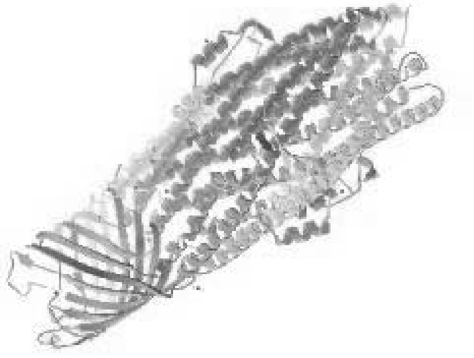

Figure 4: β-barrel structure of TolC outer membrane protein of *E. Coli*.

1.2 Transmembrane protein topology

Tagging and gene fusion are the two major approaches used biochemically for exploring TM protein topology. In general, there are four kinds of topologies: N_{in}-C_{in}, N_{out}-C_{in}, N_{in}-C_{out} and N_{out}-C_{out}.

The N_{in}-C_{in} topology is where both N- and C-terminus reside on the cytoplasmic side of the TM protein. The N_{out}-C_{in} topology is where N-terminus is on the exoplasmic side, whereas C-terminus is on the cytoplasmic side of the TM protein. By the same line of reasoning, topology N_{in}-C_{out} and N_{out}-C_{out} can be deduced similarly.

Figure 5 illustrates a model for the topology of a hypothetical TM protein**.** Since both its N- and C-terminus are on the cytoplasmic side of the membrane, it is an example of N_{in}-C_{in} topology**.** It has six membrane-spanning regions (the helices) connected by three extracellular loops (A, C and E) and two intracellular loops (B and D). On one of the extracellular loops (loop E) there is an external functional domain (in pink) and a globular region, whereas on one of the intracellular loops (loop B) there is an internal functional domain (in green) and a globular region. The helix of a TM protein is the region that resides between the lipid bilayer, whereas sidedness is referred as either the cytoplasmic (inside) or the exoplasmic side (outside) of the TM protein. Since the lipid bilayer is hydrophobic, the helix region is more hydrophobic than the loop region of the TM protein.

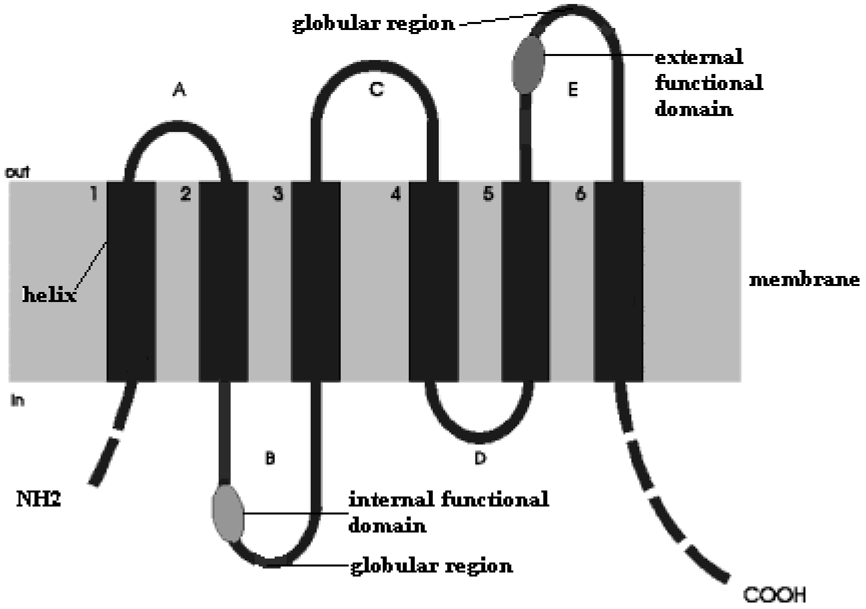

Figure 5: A model to illustrate the typical topology of a hypothetical transmembrane protein (modified from [9]).

Membrane topology is determined by how newly synthesized proteins are inserted into the membrane. This requires an understanding of TM protein assembly.

1.3 Transmembrane protein assembly

There are two major issues associated with membrane protein assembly:

First, how is each individual membrane protein targeted to its proper destination? What distinguishes a membrane protein in the plasma membrane from one in the inner mitochondrial membrane or one in the endoplasmic reticulum (ER)? This is a complex biological sorting problem. It requires distinct signals within each polypeptide as well as recognition apparatus.

Second, how are membrane proteins inserted into the membrane and how do they attain the proper topology? Do insertion and orientation also require special signals and apparatus?

1.3.1 Protein targeting

Through gene fusion, much has been elucidated about signals in the polypeptides, which direct each protein to its proper location [10]. Experiments were conducted mainly in the endoplasmic reticulum and Gram-Negative bacterium *E. coli*. According to Gennis, there are two kinds of sorting signals: primary and secondary [11].

1.3.1.1 Primary signals

Often at the amino terminus there is a recognition site or signal sequence, which directs the individual polypeptide to the target membrane (for example, membrane of nucleus, mitochondrion, chloroplast, peroxisome, and ER). This “signal hypothesis” was first postulated by Blobel *et al.* (Figure 6) [12, 41].

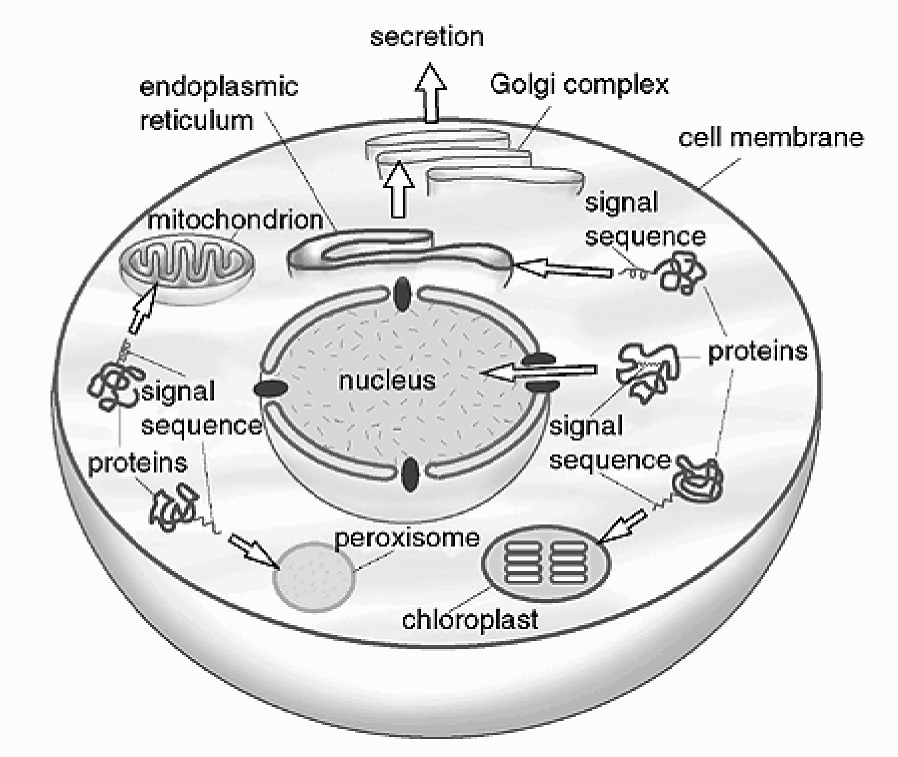

Figure 6: A schematic representation of signal sequences in directing polypeptides to the organelles, cell membrane and extracellular matrix (protein targeting).

Primary signals are highly divergent and are recognized by the translocation machinery via a specific receptor in the organelles. A signal sequence can be either a signal peptide or a signal anchor.

A signal peptide is an N-terminal peptide typically between 15 and 40 amino acids long and is not a transmembrane segment. It is hydrolyzed by a specific signal peptidase after inserting into the target membrane.

Neither the length nor the amino acid sequence is conserved for signal peptides, and mutagenesis studies have demonstrated that considerable structural variations are tolerated [11]. However, there are three structurally distinct regions in signal peptides:

(i) A positively charged amino-terminal region (n region);

(ii) A central hydrophobic core of 7 to 15 amino acids (h region);

(iii) A polar carboxyl-terminal region containing the cleavage site (c region) [11, 13].

The total hydrophobicity and length of the h-region of known eukaryotic signal peptides are intermediate between those of the most hydrophobic segments in eukaryotic cytosolic proteins and those of typical transmembrane segments.

In the illustration below, A1AT_HUMAN (Alpha-1-antitrypsin precursor with 418 amino acids) has a signal peptide. Only the N terminal 54 amino acids are shown here [Figure 7]. According to SignalP 2.0 prediction result [13], the cleavage site is between alanine (A) and glutamic acid (E) marked by a vertical bar. In figure 7 viewing from left to right, green is the n-region, cyan is the h-region and pink is the c-region.

MPSSVSWGILLLAGLCCLVPVSLA|EDPQGDAAQKTDTSHHDQDHPTFNKITPNL

Figure 7: The signal peptide of protein A1AT_HUMAN.

A signal anchor, on the other hand, is the uncleaved signal sequence that is a transmembrane segment of a TM protein. However, a signal peptide can be mistaken as a transmembrane segment by a prediction program [2] (Figure 8) and a secretory protein can be mistaken as a TM protein (Figure 9). Detailed information of signal anchors will be covered in the “insertion mechanism” section.

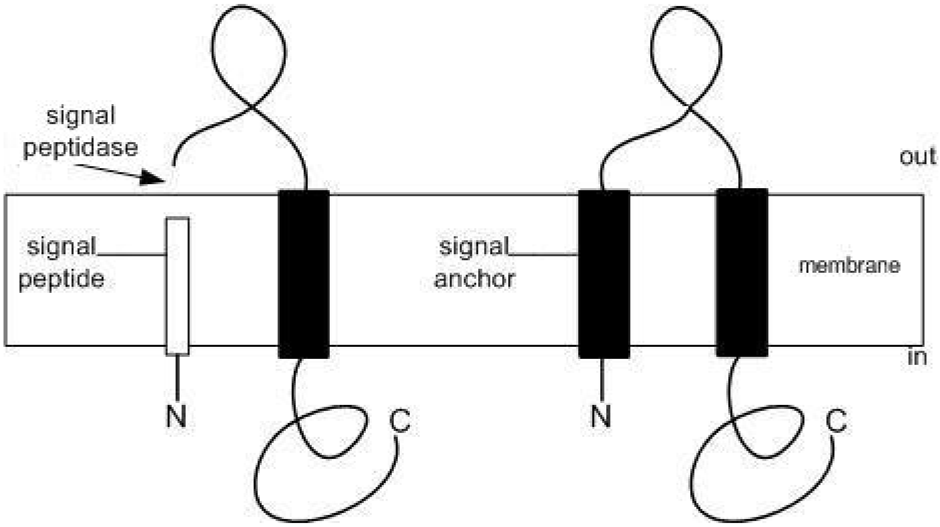

Figure 8: Illustration to show how a signal peptide can be erroneously predicted as a TM segment. The left hand side TM protein with only one TM segment can be predicted as the right hand side TM protein with two TM segments.

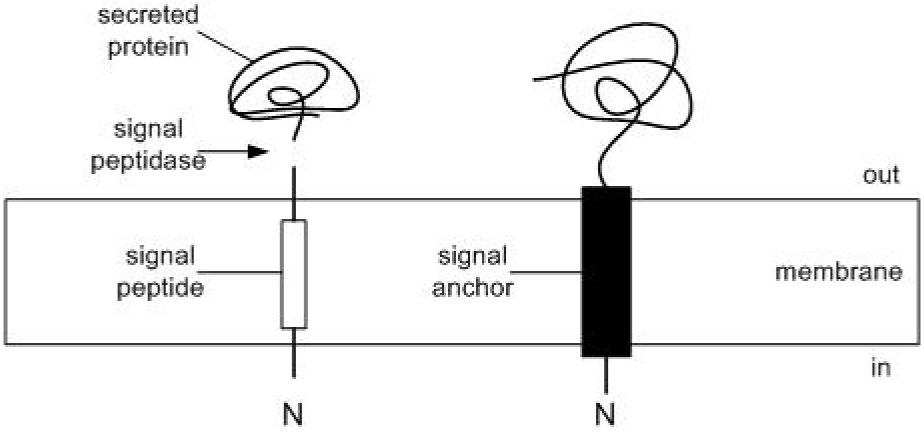

Figure 9: Illustration to show how a secretory protein can be erroneously predicted as a TM protein. The signal peptide of the secretory protein on the left hand side can be predicted as the signal anchor of the right hand side TM protein.

The signal sequence determines if the protein is secreted or remains in the membrane and the orientation of the amino terminus of the membrane protein [11].

1.3.1.2 Secondary signals

Once the proteins have become associated with the appropriate organelle, further sorting (for example, along the exocytic pathway, in mitochondrion, and in bacteria) requires additional information, which must also be encoded in each polypeptide sequence. They are the secondary signals, which determine the final destination.

1.3.2 Insertion mechanism

1.3.2.1 Start and Stop transfer segment

Signal sequences that are not removed proteolytically usually remain as transmembrane segments, or signal anchors (SA), and can initiate the translocation of flanking polypeptide on either amino (reverse SA) or carboxyl side (SA). An SA is a start transfer segment and is a hydrophobic segment, which can initiate insertion in N_{in}-C_{out} orientation, whereas a reverse SA initiates insertion in N_{out}-C_{in} orientation.

A stop transfer segment is defined as a hydrophobic segment, which halts translocation and becomes a transmembrane segment. However, it has been shown that sequences, which halt transfer in one context, can initiate transfer in another. Hence, not only the nature of the stop or start transfer sequences themselves is important, but also the surrounding polypeptide is important [11].

1.3.2.2 Insertion models

There are two biological models proposed for the insertion mechanism. The linear sequential model delineates that the hydrophobic segments insert sequentially into the membrane from N-terminus to C- terminus. Consequently, the N-terminal segment determines the orientation of the TM protein. On the other hand, the spontaneous model depicts that contiguous transmembrane segments can insert into the membrane together as a hairpin (not by hydrophobicity alone). Each model has its own supporting evidence. According to the linear sequential model, the first TM segment may decide the sidedness of membrane proteins. However, this is not always true since interactions between TM segments do exist and can cause different topologies.

Chapter 2

Computational Modeling

For most membrane proteins, especially eukaryotic, it is extremely difficult to get sufficient amount of purified membrane proteins in generating diffraction quality crystals. Therefore, *in silico* prediction tools are an important way to study TM protein structures. These tools predict the helix positions and sidedness of TM proteins from their amino acid sequences.

2.1 Features of TM proteins for *in silico* modeling

2.1.1 Hydrophobicity

According to the lipid bilayer mosaic model, hydrophobic lipid tails are oriented towards the interior of the membrane and the hydrophilic heads towards the exterior of the membrane. The core of the membrane is hydrophobic which prevents water from diffusing freely. Thus, TM segments of membrane proteins, which reside in the core of the membrane, are more hydrophobic than the parts exposed in the aqueous environment. They tend to have more hydrophobic residues such as leucine (L), isoleucine (I), and valine (V) [14]. Hydrophobic amino acids are often slightly amphipathic. Von Heijne mentioned that the length of the hydrophobic segment also decides its orientational preference. Long hydrophobic segments favor the N_{out}-C_{in} orientation, and short segments favor the N_{in}-C_{out} orientation [15].

2.1.2 Positive-inside rule

Charged residues are not symmetrically distributed. The positively charged residues arginine (R) and lysine (K) are mainly found on the cytoplasmic side of TM proteins and play a major role in determining orientation. This rule also applies to membrane proteins of intracellular organelles [7]. A strongly hydrophobic segment can be prevented from inserting into the membrane if it is flanked by positively charged residues on both ends. On the other hand, a polar segment can insert into the membrane if it is flanked by hydrophobic segments, which have the same orientation preferences [15].

2.1.3 Helix-helix interaction

It has been shown that insertion of transmembrane segments depends on neighboring segments in polytopic (multi-spanning) TM proteins in both ER and bacterial membranes.

Strong and specific interactions between α–helices of integral membrane proteins are important in their folding and oligomerization [16]. Figure 10 shows the so-called “two-stage model”. It illustrates the tertiary fold of a membrane protein from two stable transmembrane helices as a result of the helix-helix interaction [17, 51].

Proline residues occur more often in the α–helices of polytopic membrane proteins than in α–helices of soluble proteins, and often cause a kink. Transmembrane helices are often amphipathic, where the more polar surface tends to interact with other helices and prosthetic groups with the lipids [16].   However, the nature of helix-helix interaction has not been completely revealed yet.

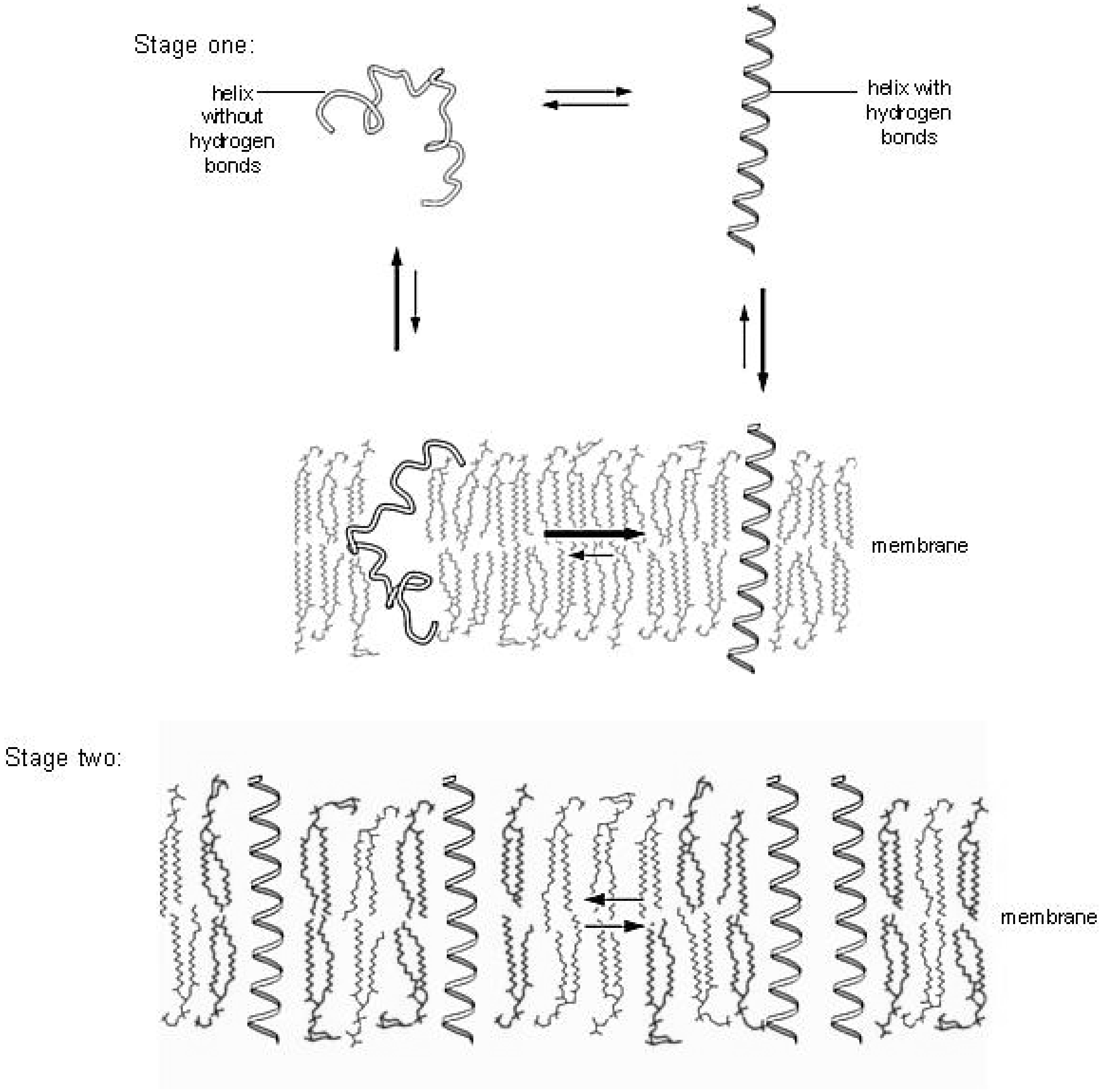

Figure 10: Two-stage model for the folding of alpha-helical integral membrane protein. The first stage is the formation of independently stable transmembrane helices resulting from hydrophobicity and the formation of main-chain hydrogen bonds in the non-aqueous environment. The second stage is the interaction of the helices to form the tertiary fold of the polypeptide.

In short, we need to model topogenic signals embedded in TM proteins. They are hydrophobicity, positive-inside rule and helix-helix interaction. In addition, cytoplasmic and exoplasmic loops must alternate.

2.2 Overview of current TM protein topology prediction programs

In general, there are two kinds of approaches to predict TM protein topology: one is local and the other is global.

2.2.1 Local approach

A few methods are based on the local approach. Some of them are as follows:

1) TopPred II [19]

This program calculates the hydrophobicity score of sliding windows to determine the helix regions and then uses the positive-inside rule to determine sidedness. Window size 19 and GES hydrophobicity scale [18] are used. Sidedness is predicted differently for prokaryotic and eukaryotic membrane proteins. For prokaryotic membrane proteins, the number of positively charged residues is counted for each side of the membrane (loop) and loops longer than 60 residues (except the first N-terminal loop) are not considered. However, for eukaryotic membrane proteins, three criteria are applied for topology prediction:

a. The difference in the number of positively charged residues between cytoplasmic and exoplasmic side of the membrane (loops);

b. The net charge difference (R, K, E, D) between the flanking 15 N-terminal and 15 C-terminal amino acid residues of the first TM segment;

c. The compositional distance [14] for loops longer than 60 residues.

Hydrophobicity and charge bias are the local predictors for TM segments and sidedness prediction [19, 20]. However, the chosen window size limits the actual length of the helix.

2) TM Finder [21]

This program is a combination of segment hydrophobicity and non-polar phase helicity scales developed from peptide studies. A candidate TM segment must satisfy both hydrophobicity and helicity thresholds. It treats false split by means of gap-joining operation. It only predicts locations of TM segments. TM Finder uses a sliding window as well [21].

3) SPLIT [22]

This program associates multiple scales of amino acid attributes with secondary structure conformations for each amino acid. One of such scale is the sequence hydrophobic environment. The sequence hydrophobic environment of an amino acid is the average hydrophobicity of its five left and five right neighbor amino acids in the protein. The secondary structure conformations are α-helix, β-sheet, turn (4 amino acids on each side of the helix) and undefined conformation. Prediction is made by comparison of preferences for each residue in the sequence. Its preference function was based on the Kyte-Doolittle hydrophobicity scale [50], but as a nonlinear function of the sequence hydrophobic environment. It uses sliding window implicitly and sets up filter parameters to prevent false merge and to distinguish normal length TM helices, short TM helices (13-16 residues long and α-helix preference above certain threshold), and membrane-buried helices. Membrane-buried helices are not counted as TM segments. In addition, SPLIT predicts helix positions only [22].

2.2.2 Combined method

The combined method is a combination of both local and global methods. One example is the Consensus Predictions. It uses 5 methods, including TMHMM, HMMTOP, MEMSAT, TOPPRED and PHD by a simple majority-vote approach [23]. However, some of the proteins used in the test set have been used for training by those 5 methods.

2.2.3 PHDhtm [27]

PHDhtm is a Neural Network (NN) based method for TM protein topology prediction. Originally, PHDhtm was only used for helix location prediction. Later on, it incorporated the positive inside rule to predict sidedness as well. The input is from a multiple sequence alignment profile. However, this method uses a certain size window (local approach) to train the net that has a feed-forward topology. However, sometimes helices predicted by PHDhtm alone were too long. PHDhtm could hardly distinguish the loop between two helices if the loop is fairly hydrophobic. In order to overcome the weakness inherent to the method (i.e. false merge), a dynamic programming method, global approach, was later introduced into PHDhtm to further verify the prediction result, which is termed PHDhtm_ref [24, 25, 26, 27].

However, the NN implementation has the following weaknesses: its topology is not biologically intuitive compared to a hidden Markov model, and the model may only represent the training result from a local optimum. In addition, it is a black box approach. In other words, even if the prediction accuracy is fairly high, we would not be able to know the underlying mechanism.

To summarize, the main weakness of the local approach is the lack of specificity. On the other hand, global approach examines sequences as a whole and does not set any empirical cutoffs and rules. Moller *et al*. did an experiment on a set of 87 membrane proteins and the prediction accuracy for TopPredII, PHD, Memsat 1.5, HMMTOP and TMHMM 1.0 and TMHMM 2.0 (for details please see the next subsection) were 7%, 20%, 38%, 22%, 38% and 41% respectively [1]. TM Finder and SPLIT predict helix positions only. Below we will take a close look at hidden Markov model, a global approach used in TM protein topology prediction.

2.2.4 HMM [28]

A hidden Markov model (HMM) is a statistical, probabilistic, and generative model. It is a doubly embedded stochastic process. One is hidden and the other is observable. At any time, only the sequence of output symbols is observed, but the states that emit the output remain hidden. There are exponentially many state paths π corresponding to a given sequence x. The probability of observing a sequence x is therefore $P(x)=\sum_{i} P(x,\pi_i)$, where i is the index of state paths. However, through either the Forward or Backward algorithm [28], $P(x)$ can be calculated in a quadratic order of the number of states. That is $P(x_1 ... x_i, S_i = k)$ or $P(x_{i+1} ... x_L, S_i = k)$, where S_i is the state at position i, which is k; and i is the index of the sequence x from 1 to L. The probability for each state at position i depends on the probabilities of previous incoming states, the transition probabilities from previous state to current state and the emission probability of current state to emit x_i.

We use Viterbi algorithm [28] to find the most probable state path π_i to be the optimal state path for a given sequence, i.e. the state path that maximizes $P(x,\pi)$. Given any finite sequences as training data, there is no optimal way to estimate the model parameters [46]. However, with state path unknown, we use Baum-Welch algorithm to locally maximize $P(x \mid \theta)$ over all the training sequences. That is,

$$a_{kl} = \frac{\sum_{j} \sum_{i=1}^{L-1} P(S_i = k, S_{i+1} = l \mid x^j, \theta)}{\sum_{j} \sum_{i=1}^{L-1} P(S_i = k \mid x^j, \theta)}, e_k(b) = \frac{\sum_{j} \sum_{i=1}^{L} P(S_i = k, x_i^j = b \mid x^j, \theta)}{\sum_{j} \sum_{i=1}^{L} P(S_i = k \mid x^j, \theta)}$$, where a_{kl} is the reestimate of transition probability from state k to state l over j training sequences; $e_k(b)$ is the reestimate of emission probability for state k to emit symbol b over j training sequences; and θ is the current set of model parameters. For modeling labeled sequences, only valid paths are counted. They are paths whose state labels are the same as the sequence labels [28]. HMM is biologically intuitive and can model both symbol (i.e. amino acid) distribution and length distribution (i.e. loop and helix).

However, the weakness of HMM is that the model parameters (transition and emission probabilities) obtained from training might only be local maxima. Besides, HMMs do not model distant dependency well. The first order HMMs (above) at position i only depends on previous states at position $i-1$. Even for higher order HMMs, they could only model a limited and fixed number of dependencies but with much higher complexity. An *n*th order Markov chain over an alphabet Σ is equivalent to a first order Markov chain over the alphabet Σ of *n*-tuples [28]. The other limitation is the assumption that successive symbols are independent. Therefore, the probability of a given sequence can be written as a product of probabilities of each individual symbol. This is apparently not true in TM protein prediction. For example, if an amino acid is inside of the membrane, then the next amino acid has certain probability of being inside or in helix, but cannot be outside of the membrane. Training for HMM hinges on the hope to obtain all signals (hydrophobicity, positive inside rule, etc.) that could be used for prediction. However, not all signals could be obtained especially with less abundant known topology sequences for training. A summary of HMM approaches in addressing TM protein topology prediction is provided below.

2.2.4.1 Membrane protein structure and topology (Memsat) [7]

The model contains five structural states: inside loop, outside loop, inside helix end, helix middle and outside helix end. It uses a dynamic programming algorithm to determine the optimal location and orientation of a given number of TM helices. The highest scoring number of TM helices is selected as the best prediction. It uses separate propensity scales (equivalent to emission probabilities of a HMM) for residues in the cap (helix end) and helix core region of the membrane. They are set to be 4 and 17-25 residues respectively [7] (Figure 11).

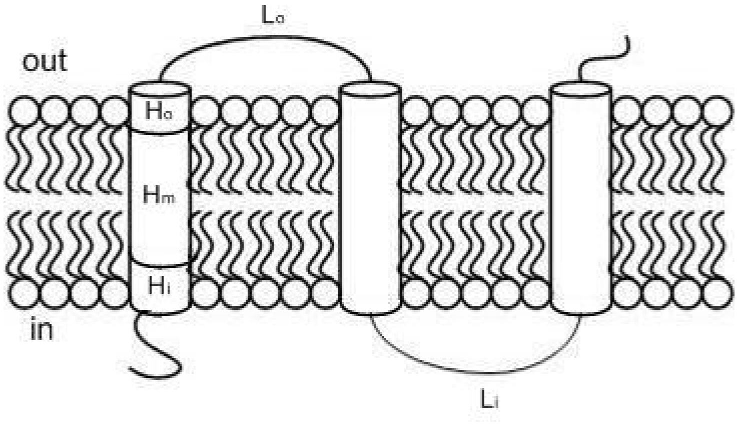

Figure 11: Structural states defined by Memsat for a typical helical TM protein (reproduced from [7]).

2.2.4.2 HMM for topology prediction (HMMTOP) [29]

This model also contains five states: inside loop, inside tail, membrane helix, outside tail and outside loop. Tails are thought to interact with the heads of lipid bilayer, while loops do not. Two tails between helices form a short loop, but tail-loop-tail between helices form a long loop. This model topology is similar to Memsat. The differences are in the localization and interpretation of helix tails, which were called helix ends in Memsat. Helix tails are not in the membrane, whereas helix ends (the very ends of helices) are in the membrane.

Short loops with lengths between 5 and 30 amino acid residues were found significantly more often than expected (a different distribution than geometric distribution) by Tusnady and Simon [29]. A geometric distribution is the background (neutral) amino acid distribution. It can be represented by a self-looping state. In response, they implemented the length distribution of short loops as well as helices. The length of a helix is 17-25 residues and the length of a tail is 1-15 residues. The prediction accuracy (the number of protein topologies predicted correctly compared with the annotated ones) is better than Memsat [29]. The implementation of HMMTOP is similar to TMHMM on helix and loop structure.

2.2.4.3 Transmembrane HMM (TMHMM) [2, 31]

This model contains seven different states: one for the helix core, two for caps on either side, one for loops on the cytoplasmic side, two for short and long loops on the non-cytoplasmic side, and one for ‘globular domains’ in the middle of each loop. It is postulated that seven states may be more sensitive to the variation of the amino acid compositions than five states [30]. For each distinct state, there are a number of states joined with its emission probability (Figure 12). TMHMM is a constrained HMM (For each state, transitions are among a limited number of states, not for all states). Thus, the transition matrix is a sparse matrix. Tied states are due to the limited number of known topology protein sequences to train from, to avoid overfitting. Technically, there is no difference between TMHMM 1.0 and TMHMM 2.0 except that TMHMM 2.0 was retrained on the same data set [1].

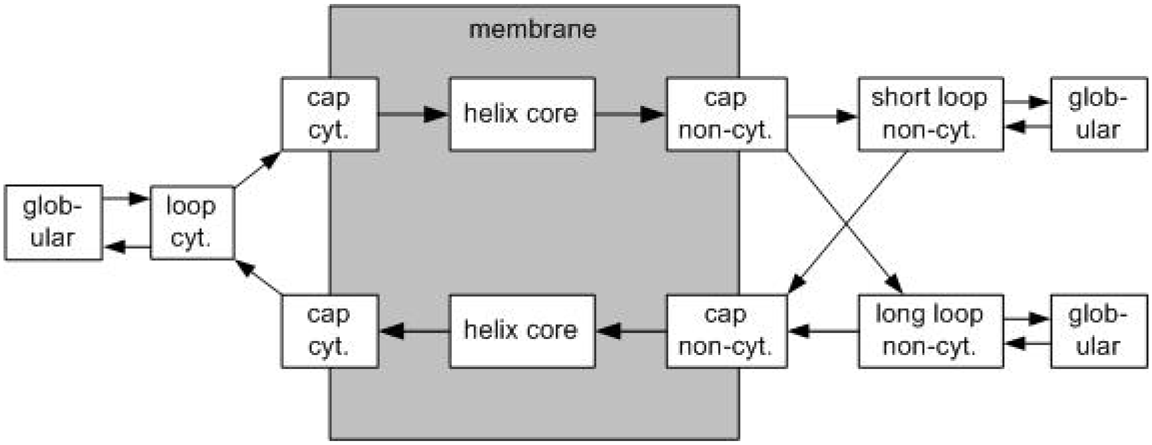

Figure 12: The overall layout of TMHMM. Each box corresponds to one or more states in the HMM. Parts of the model with the same text are tied, i.e. their parameters are the same. Cyt. represents the cytoplasmic side of the membrane and non-cyt. stands for the exoplasmic side (reproduced from [31]).

2.2.4.3.1 Length of helix cap region and helix

Helix and loop lengths are two constraints. Both Memsat and TMHMM embody the belief that the head region of the lipid bilayer contains many polar and charged residues and makes contact with the phosphate groups of the lipids. Thus, they model it as two ends of helices. However, HMMTOP models it as tails, which reside outside of helices.

The length of this region is arbitrarily taken as 4 in Memsat. Sonnhammer *et al*. further discovered that accuracy dropped significantly with caps less than 4 residues, while caps of 4-7 residues rendered the same result. They picked 5 (Figure 13) and modeled the helix core region of 5-25 residues long. This allows the length of helices to be 15-35 residues long, whose range is the longest among the three models (Figure 14).

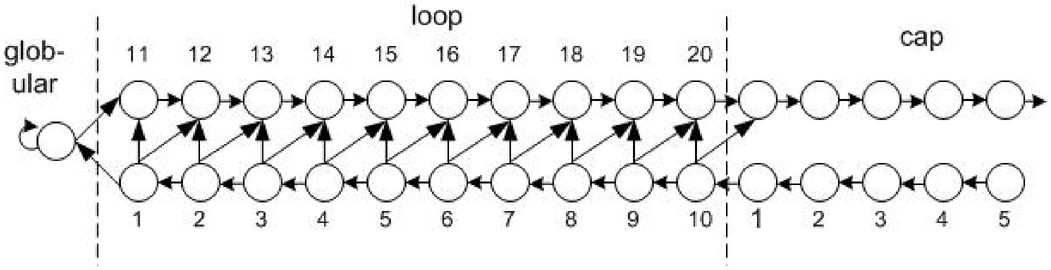

Figure 13: The detailed structure of the loop and helix cap models in TMHMM (reproduced from [31]).

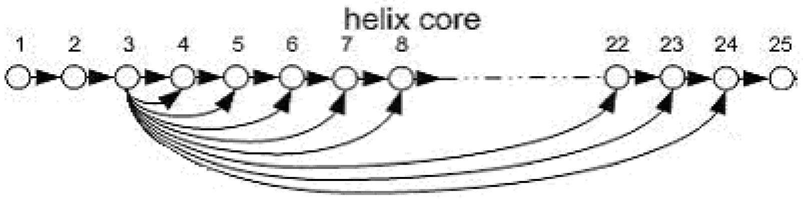

Figure 14: The detailed structure of the helix core model from TMHMM, which models lengths from 5 to 25 residues long (reproduced from [31]).

2.2.4.3.2 Loop architecture

Sonnhammer *et al*. claim that the difficulty in predicting the topology seems to partly arise from the fact that substantial number of positively charged residues in the globular domains of non-cytoplasmic side loops blurs the positive-inside rule. In bacteria, positively charged residues in different length of loops do not show the same effect [32]. Positively charged residues in short loops can prevent helices from translocation across the membrane. However, positively charged residues in long loops do not necessarily halt the translocation; instead they may be translocated across the membrane.

When training the ‘short’ path on loops shorter than 100 residues and the ‘large globular domain’ path on longer loops on the exoplasmic side, the accuracy increased by 6-14%. However, having two alternative loop paths on the cytoplasmic side reduced the accuracy by 2-11%. The highest accuracy was observed at loop ladder length (the length for a loop before and after entering into the globular region) between 2x10 and 2x15. Based on these observations, they used two loop paths on the non-cytoplasmic side to model short and long loop respectively and only one on the cytoplasmic side. Besides, the loop ladder length is 10 amino acids long [31].

So far, the prediction accuracy is fairly low even with the current best global approaches. In addition, sidedness prediction accuracy is lower than helix prediction accuracy. Furthermore, HMMs cannot model helix-helix interaction. We propose the following possible approaches for further improvements.

2.3 Potential improvements

Evaluation of current TM protein topology prediction programs has demonstrated the need to improve prediction accuracy. In an attempt to improve the current best prediction program, TMHMM, we propose three possible approaches:

2.3.1 Incorporation of cytoplasmic-specific and exoplasmic-specific functional domains into TMHMM to improve the prediction accuracy

This approach could be illustrated by Figure 15 and Figure 16. TMHMM might generate a wrong topology for a putative TM protein (upper diagrams of Figure 15 and Figure 16). However, if we know that this protein has an internal domain (which appears preferentially inside of the membrane), we may then boost its probability of being inside, and thus yield the correct prediction (lower diagrams of Figure 15 and Figure 16).

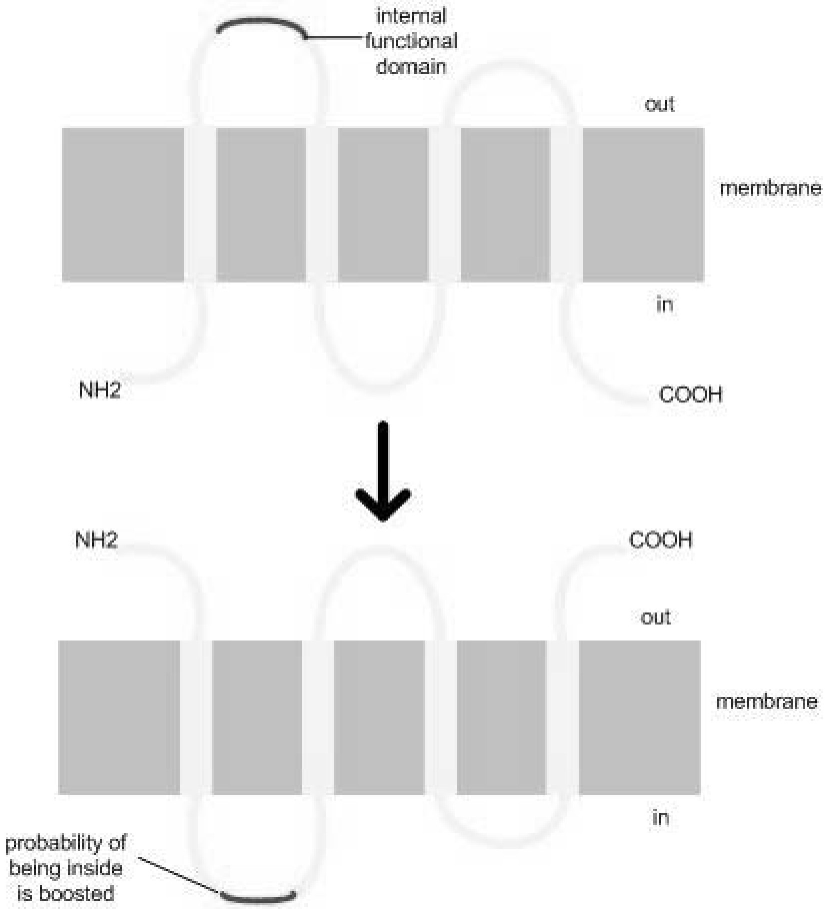

Figure 15: Graphical illustration to show how a sidedness error can be corrected through the external incorporation of functional domains into TMHMM.

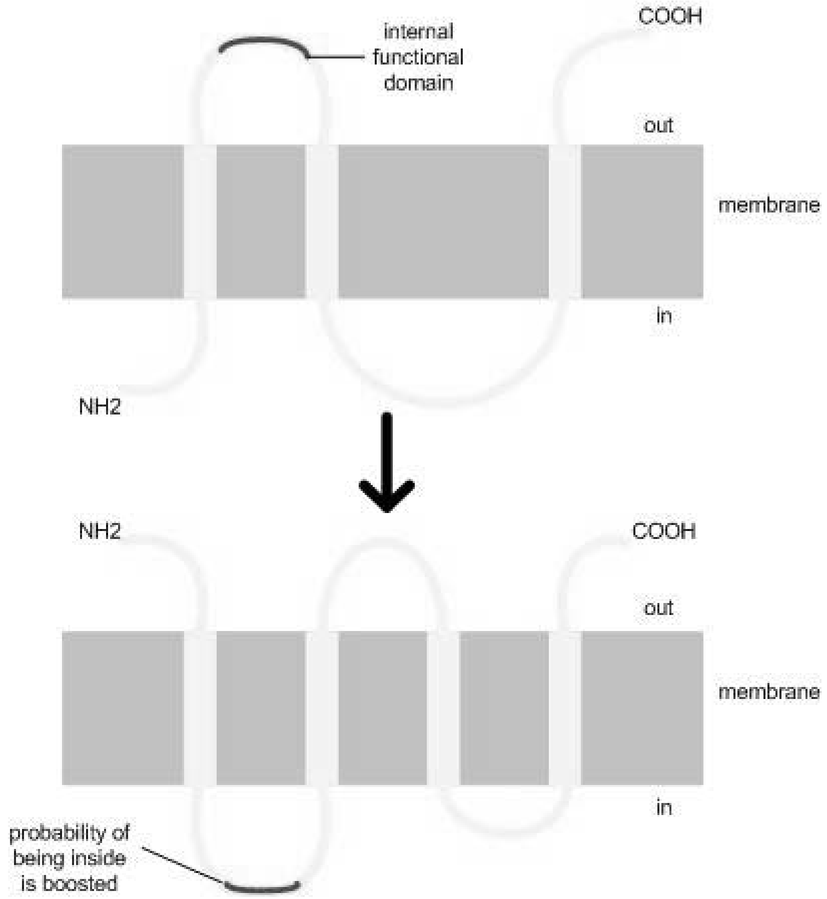

Figure 16: Graphical illustration to show how the overall TM topology prediction (helix number + sidedness) can be improved through the external incorporation of functional domains into TMHMM.

HMMTOP 2.0 added some preliminary experimental information (including pattern predictors) on top of the HMMTOP 1.0 to help improve prediction accuracy. It allows the user to localize one or more sequence segments in any of the five structural regions used in HMMTOP. For example, proteins in ABC (ATP Binding Cassette) protein family contain three cytoplasmic motifs, the Walker A, B and the ABC-signature sequence motif. With the help of these cytoplasmic motifs, HMMTOP 2.0 could correctly predict the topology of the MRP1 protein. However, this information has to be given by the user.

Moller *et al.* also suggested using additional information such as protein domains or post-translational modifications when the prediction from TMHMM is in doubt [1]. However, information on protein domains or post-translational modifications has not been explicitly implemented into any of the programs.

Generally speaking, there are two ways to incorporate pattern predictors. One is to incorporate them externally into the HMM, while the other is to incorporate them internally. The former is to adjust the probabilities of certain topologies at the position of the predictor. The latter is to hardcode the pattern information into the HMM structure, for example, the loop region of TMHMM.

External incorporation boosts the probability of the topologies, which predict internal domains as internal and/or external domains as external and decreases the probability of other topologies accordingly. In other words, the external incorporation is built on top of the TMHMM. However, the internal incorporation has to make changes in the model (i.e. the transition probability matrix). For example, we can train an HMM for each pattern or domain and add it into the loop region of TMHMM. The transition probabilities between the loop and pattern or domain have to be trained after the addition of pattern and/or domain HMMs (Figure 17).

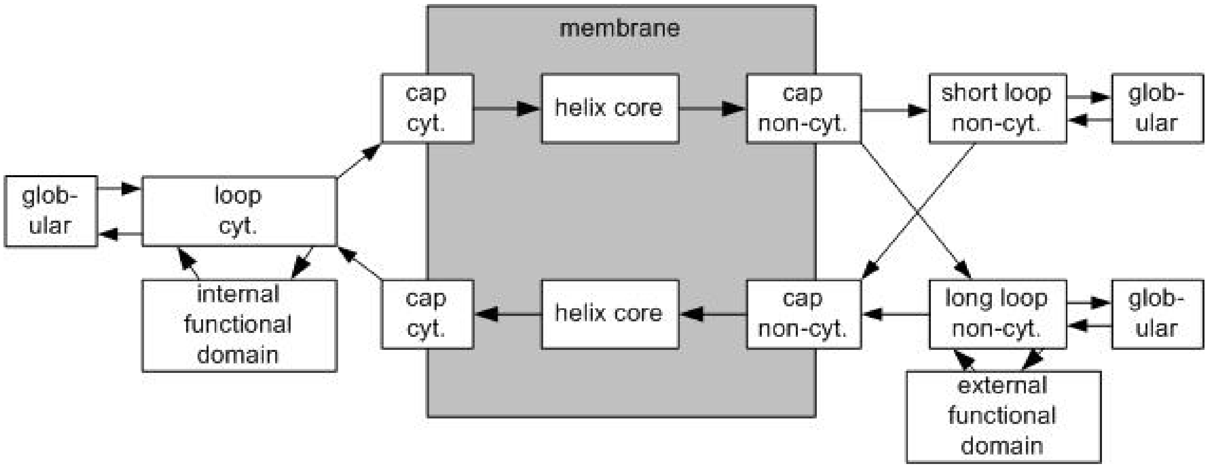

Figure 17: Illustration to show an internal incorporation of pattern and domain information into TMHMM.

One drawback with the internal incorporation is that internal changes and training have to be made every time if there is a change on the functional domains. The worst-case scenario could be state space explosion if transitions are made on condition of different combinations of features (i.e. pattern and domain information). For example, if there are 6 patterns and domains, and transition probability $P(S \mid x, F_1, F_2, F_3, F_4, F_5, F_6)$ (S is the current state, x is the observation, F_i represents one of the pattern or domain, $i = 1 \ldots 6$, F_1 to F_6 are the combination of 6 patterns and domains) depends on the combinations of the 6 patterns and domains, then there are 2^6 combinations of the patterns and domains. One state now becomes 2^6 different states. If there are more pattern and domain combinations, a state space explosion could result. Thus, we used external incorporation to improve prediction accuracy of TMHMM.

2.3.2 Generation of a merged HMM from different classes of TM protein HMMs guided by a traffic cop

In this method, classes of TM proteins can be differentiated by their sequences. For example, G-protein coupled receptors (GPCRs) or ion channels. GPCRs generally have seven TM segments with an extracellular N-terminus. Ion channels have at least four TM segments. We can train a version of HMM on each of these classes separately. Then, create a merged HMM with a traffic cop that routes sequences to the appropriate HMM for analysis and see if this merged HMM performs better than the original TMHMM (Figure 18).

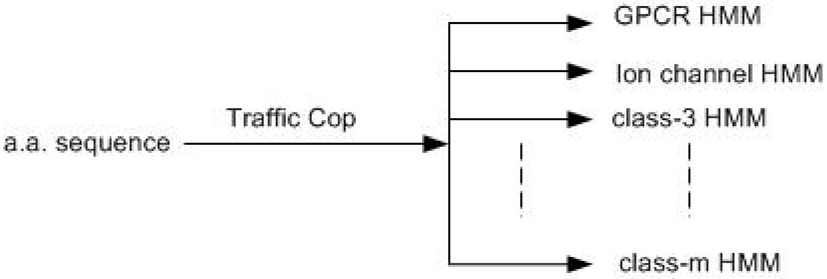

Figure 18: The merged HMM layout in which a traffic cop routes the query sequence to the appropriate class of HMM for analysis.

Because the use of this method requires a good representative model for each class as well as a good traffic cop to differentiate between each class of proteins, this method is rather limited by its specificity and lack of generality.

2.3.3 Model interactions between transmembrane helices

The insertion study of membrane proteins [10] points out that, in addition to hydrophobicity, the orientation of a segment towards the membrane, the presence of up- and downstream sequences, and the interactions between different topogenic signals also play a significant role during insertion. Interactions between different topogenic signals can result in the exclusion of a hydrophobic segment from the membrane or the insertion of a less hydrophobic sequence into the membrane. This has not been modeled by any of the current prediction programs on TM proteins. However, the nature of helix-helix interaction has not been totally characterized yet.

Much experimental evidence indicates that in addition to start and stop transfer segments, other signals also affect the TM protein topology. The following illustrations show how insertion depends on the downstream segments (Figure 19) from the studies on the anion exchanger Band 3 (A), the protein translocation complex subunit Sec61 (A and B) and the citrate transporter CitS (C).

A. A moderately hydrophobic segment at the cytoplasmic side may be inserted into the membrane if a downstream hydrophobic segment exists and has strong stop transfer ability (the “driving” segment). They can be inserted either spontaneously as helical hairpin or assisted by chaperone-like proteins (Figure 19A).

B. A moderately hydrophobic periplasmic or luminal segment may be inserted if a downstream hydrophobic segment exists (Figure 19B).

Example A and B show that segments with weak insertion signals might be inserted into the membrane if their immediate downstream segments exist.

C. The insertion machinery does not translocate the driving segment across the membrane until the less hydrophobic segment is also exported (in Figure 19C segment VIII of CitS is prevented from insertion until segment IX is translated and translocated to the periplasm).

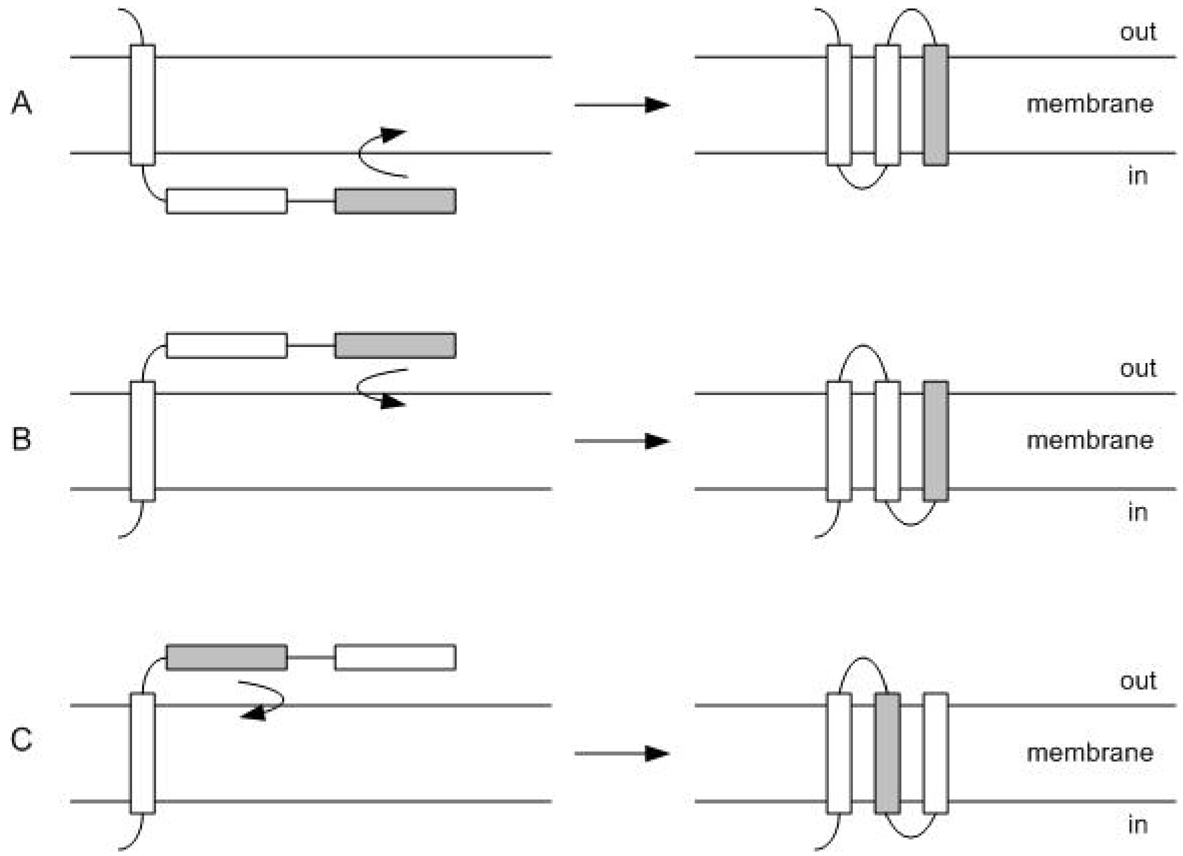

Figure 19: Hypothetical insertion intermediates. The insertion intermediates (left), two consecutive TM segments have not been inserted into the membrane. The hydrophobic segment, in grey rectangle, drives the insertion of its preceding (A and B) or following (C) segment into the membrane (right) (reproduced from [10]).

The reason we chose to implement and assess the first approach is simply because

1) Conceptually this requires the minimum change to HMM;

2) Functional domain databases (e.g. PROSITE, Pfam, Smart etc.) have become comprehensive in recent years;

3) Compared to the first approach, the second approach is less general and it largely depends on the ability to differentiate between different classes of proteins;

4) The third approach may be a plausible solution, but it requires a more developed model of TM proteins.

Originally our attempt was to incorporate internal and external functional domains to improve sidedness prediction (by the method ‘fixed helix HMM’). However, we discovered later that prediction on sidedness and helix position probably are not two independent issues. Since by incorporating internal and external functional domains into TMHMM (which is the augmented HMM or AHMM), TM protein topology prediction in general is improved.

Chapter 3

Experiments and Results

The central hypothesis of this thesis is that incorporation of internal and external functional domains can augment prediction accuracy of TMHMM.

3.1 Measurements

The following subsections will introduce how we selected potential functional domain predictors, how we implemented them externally into TMHMM and how we tested the robustness of AHMM. Upon selection of potential functional domain predictors, we used precision (true hits / (true hits + false positives)) and recall (true hits / (true hits + false negatives)) from PROSITE as a reference for the functional domains. The precision is equivalent to specificity whereas the recall is equivalent to sensitivity.

We used sensitivity and specificity to compare TMHMM and AHMM upon prediction on helix and sidedness. We define sensitivity as true positives / (true positives + false negatives) (the number of correct predictions out of the reference number) and specificity as true positives / (true positives + false positives) (the number of correct predictions out of the total number of predictions).

3.2 Data sets

Basically we used two sets of data for our experiment. One is the 160 protein data set from the TMHMM training set [31] and the other is the 62 data set from Moller *et al*. collection.

The160 protein data set is used for extracting internal and external functional domains. The 160 protein data set consists of 108 multi-spanning and 52 single-spanning proteins. It does not contain proteins that had yielded different topologies from different experiments and with no justification.

The data set includes both eukaryotic, prokaryotic and organelle TM proteins. They are chosen as training data to extract potential pattern and domain predictors for our augmented HMM (AHMM).

The test data is from Moller *et al*. collection. However, we excluded ER, mitochondria and all membrane proteins that have not been completely annotated and those present in either the 160 or 83[1] data set. Thus, only 62 protein sequences were used as test data. All these proteins are TM proteins with experimentally known topology.

The prediction accuracy (the percentage of correctly predicted number of sequences out of the total number of sequences for prediction) of TMHMM on the160 data set is approximately 79% whereas on the 62 data set is approximately 52%. That is, 33 out of 62 sequences were predicted correctly.

3.3 Method

We implemented two methods to improve TMHMM prediction accuracy. One is our augmented HMM, AHMM and the other we called the “fixed helix HMM”.

3.3.1 Reconstruction of TMHMM 1.0

To compare our AHMM with TMHMM, we reconstructed TMHMM by using the transition and emission probabilities of TMHMM 1.0 since only the parameters of TMHMM 1.0 are available. Changes were made on certain transition probabilities (for details please see the next subsection). However, the prediction result was compared with both TMHMM 1.0 and TMHMM 2.0. The original TMHMM used the "N- or one- best algorithm" for prediction whereas our reconstructed

[1] The 83 data set contains 38 multi-spanning and 45 single-spanning proteins, which was originally compiled by Jones *et al.* [7] and provided by Rost *et al.* [27].

TMHMM used Viterbi algorithm instead. Krogh claimed that "N or one best algorithm" was no worse than Viterbi [33].

3.3.2 Pre-experiment test

The transition and emission probabilities of TMHMM 1.0 cannot be used directly as initial parameters for training because 6 sequences from the data set of 160 proteins (surprisingly enough) and 5 sequences from the 62 data set could not be accepted (probability is zero). This is due to the following bugs within the transition probabilities provided by TMHMM:

1) Two initial transition probabilities from TMHMM 1.0 (from both short and long outside loops with length 1 to the membrane) were set up to be 0.

2) The length of outside cap (from cap to outside loops) is 4 instead of 5, which was not modeled by TMHMM 1.0.

Therefore, TMHMM 1.0 did not model:

1) Transitions within helices (originally four transitions were 0 and were changed to 1 before training).

2) Loop ladder structure for long loop and loop with length 1 into membrane (it is shown in Figure 11 TMHMM loop architecture, but was not implemented by the transition probability matrix of TMHMM 1.0).

3) Shorter cap length.

3.3.3 Details of the two methods implemented to improve TMHMM

3.3.3.1 AHMM

We have changed the way TMHMM computes the Viterbi probability of the possible topologies of an input sequence by taking advantage of signature and domain predictors found in the sequence. For example, we boost the probability of the topologies, which predict internal functional domains as internal and/or external functional domains as external to the membrane and decrease the probability of other topologies accordingly.

For a signature, the probability of topologies is modified only at its start position. For a domain, the probabilities of topologies are modified at both the start position and end position of the domain.

Our augmented model uses GenomeScan [34] technique by modifying the HMM probabilities when a signature or domain predictor is encountered. Specifically:

$$P(\pi_i, x \mid H) = \begin{cases} \left(\dfrac{P_H}{P(\Phi_H)} + (1 - P_H) \right) \bullet P(\pi_i, x), & \text{if } \pi_i \in \Phi_H \\ (1 - P_H) \bullet P(\pi_i, x), & \text{if } \pi_i \notin \Phi_H \end{cases}$$

For example, for an internal signature:

H −− the signature is internal.

$P(\pi_i, x \mid H)$ −− for sequence *x*, the probability of topology π_i at the position of the signature given that it is internal.

P_H −− the probability that the signature is internal.

Φ_H −− the set of topologies that identify the protein as internal at the position of the signature.

$P(\Phi_H)--$unaugmented probability that the site is predicted to be internal at the position of the signature.

$P(\pi_i, x)--$ for sequence *x*, the probability of topology π_i, as calculated by Viterbi algorithm for decoding HMM sequences.

$\frac{P_H}{P(\Phi_H)}+(1-P_H)$ is always greater than 1 and $(1-P_H)$ is always less than 1.

For example, from position 240 to position 440 of sequence ENVZ_ECOLI, there exists a HIS_KIN domain. It is supposed to be internal. Since TMHMM predicts this region as external, it gives the wrong prediction. However, AHMM boosts the probability for topologies being internal at both position 240 and 440 by using the first part of the formula ($\left(\frac{P_H}{P(\Phi_H)}+(1-P_H)\right)\bullet P(\pi_i, x)$, if $\pi_i \in \Phi_H$). On the other hand, it lowers the probability for topologies being external at the two positions by using the second part of the formula ($(1-P_H)\bullet P(\pi_i, x)$, if $\pi_i \notin \Phi_H$). It gives the correct prediction (Figure 20).

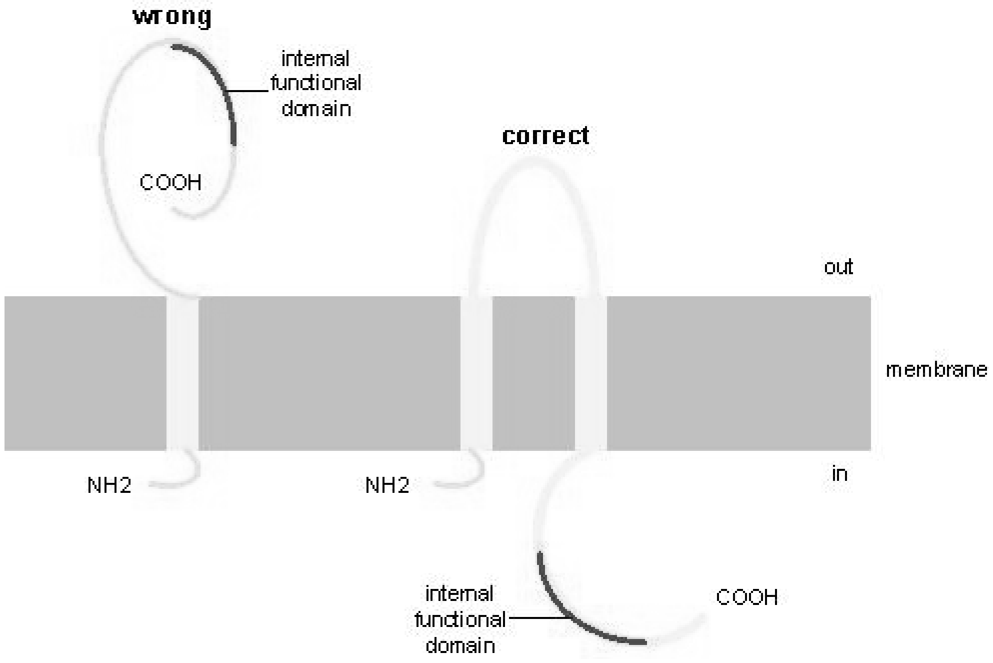

Figure 20: Topologies of ENVZ_ECOLI predicted by TMHMM (left diagram) and AHMM (right diagram) respectively.

3.3.3.2 Fixed helix HMM

In addition to AHMM, we also implemented a method called fixed helix HMM. It simply flips the sides of the TM protein topology when it detects an internal domain appears outside and vice versa after TMHMM prediction. The principle is shown in Figure 21.

When there are multiple domains in the protein, and as long as internal and external functional domains are alternating, we could still consider flip the sides of the topology. But if they generate conflicting information such as the presence of internal domains adjacent to each other instead of alternating (as shown in Figure 22), we must decide to flip or not to flip.

The conflict arises because of the wrong helix number from the original TMHMM prediction and/or wrong functional domain information. In addition, the more conflicts there exist, the less confident it is to flip. We may consider flipping by ignoring the less confident one if the rest of the domains (either internal or external ones) are consistent with each other. However, the absolute confidence of a functional domain, which is the probability of being internal or external to the membrane, is unknown most of the time. More studies could be done regarding when to flip, however, it may not be worth pursuing given the better performance of AHMM. In the case where conflict occurs, we decide not to flip and keep whatever TMHMM predicts.

To summarize, fixed helix HMM only helps in correcting sidedness errors, but not helix number errors.

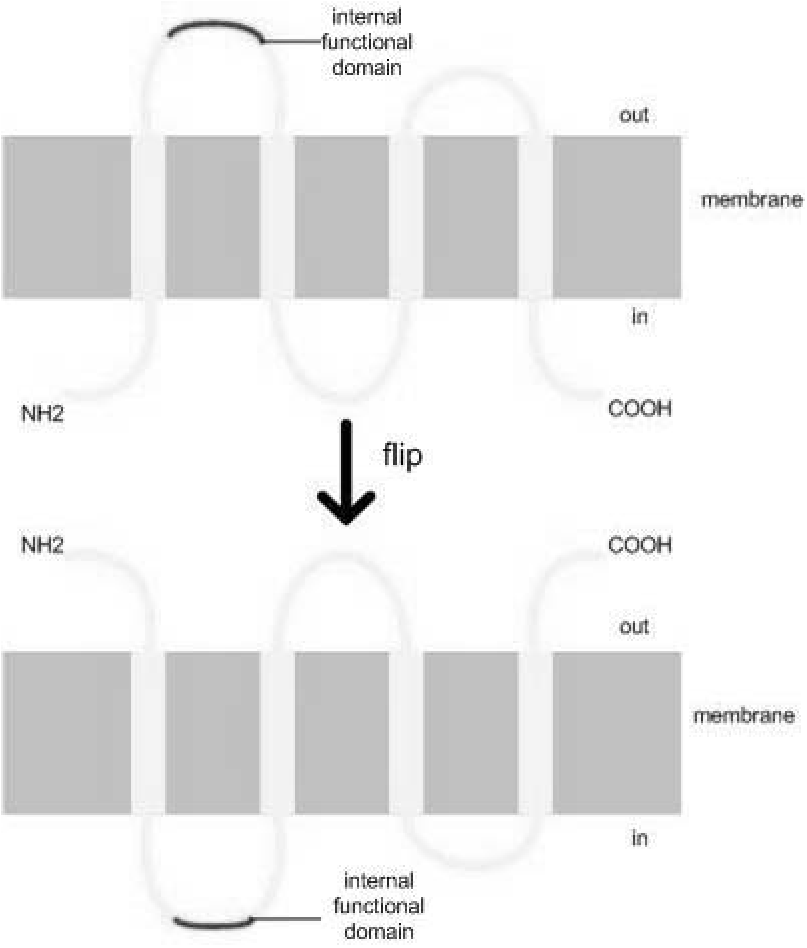

Figure 21: Illustration of how ‘fixed helix HMM’ in correcting sidedness error of TMHMM is based on functional domains.

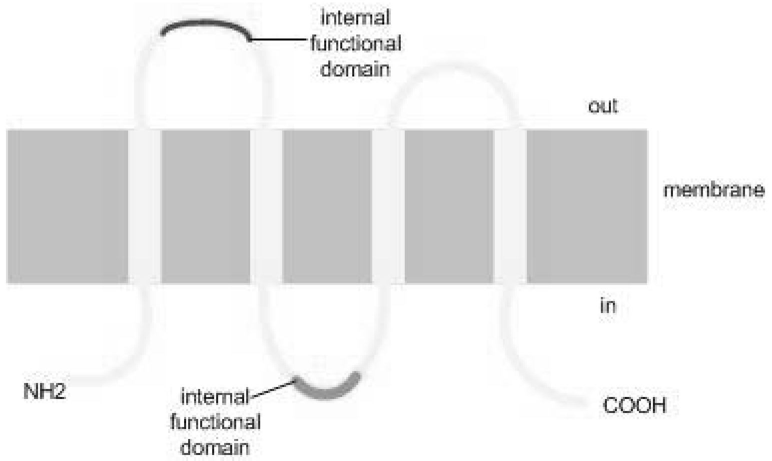

Figure 22: Conflict information of functional domains.

3.4 Definition of pattern and domain predictors

PROSITE is a method of determining the function of uncharacterized proteins translated from genomic or cDNA sequences [35].

A particular cluster of residue types of a sequence is known as a pattern, motif, signature, or fingerprint. It represents a conserved region of proteins.

In this paper, we use “signature” to emphasize a PROSITE specific pattern versus its consensus pattern. Domains refer to functional or structural domains that cannot be detected using patterns due to their extreme sequence divergence. Domains are implemented by position specific score matrix (PSSM, also known as profiles).

3.5 Selection of pattern and domain predictors

3.5.1 Biological approaches

The following assumptions are made:

1) phosphorylation sites are more likely to appear on the cytoplasmic side of a TM protein;

2) glycosylation sites are more likely to appear on the non-cytoplasmic side of a TM protein.

3.5.1.1 Two sources for phosphorylation and glycosylation motifs: PROSITE release 17.4 and NetOGlyc_2.0/NetPhos 2.0.

3.5.1.1.1 PROSITE phosphorylation and glycosylation consensus patterns:

From PROSITE release 17.4, keyword search for phosphorylation and glycosylation returned 8 phosphorylation and 1 glycosylation consensus patterns, giving a total of 9 patterns.

However, 4 phosphorylation patterns were eliminated since 1 had only one hit with our training data and the other 3 had no hit at all. This left only 5 consensus patterns in this experiment (4 phosphorylation and 1 glycosylation consensus patterns). However, the prediction results for AHMM incorporated with them were worse than TMHMM. There were more wrong predictions than right ones. We then examined the number of times that the consensus patterns occurred inside, outside of the membrane or in helix and normalized them by the total number of amino acids inside, outside or in helix of all the proteins in the set respectively (Table 1).

Table 1: Examination of the frequency of each consensus pattern being inside vs. outside (normalized).

consensus pattern	for 160			for 57[2]		
	in	out	helix	in	out	helix
C1: [RK]-[RK]-x-[ST] (internal) cAMP- and cGMP-dependent protein kinase phosphorylation site	0.0028	0.0011	6.6 E-5	**0.0041**	**0.0072**	0.0012
C2: [ST]-x-[RK] (internal) Protein kinase C phosphorylation site	0.017	0.013	0.0011	0.014	0.013	7.1 E-4
C3: [ST]-x-x-[DE] (internal) Casein kinase II phosphorylation site	0.016	0.014	9.9 E-5	**0.015**	**0.016**	0.0018
C4: [RK]-x-x-[DE]-x-x-x-Y (internal) Tyrosine kinase phosphorylation site	4.2 E-4	4.2 E-4	0	**2.4 E-4**	**3.7 E-4**	0
C5: [RK]-x-x-x-[DE]-x-x-Y (internal) Tyrosine kinase phosphorylation site	5.4 E-4	4.2 E-4	0	**2.4 E-4**	**2.5 E-4**	0
C10: N-{P}-[ST]-{P} (external) N-glycosylation site	0.0049	0.010	0.0011	0.0041	0.0072	0.0012

Note:

1. Consensus pattern C4 and C5 are both Tyrosine kinase phosphorylation site.

2. The regular expression for PROSITE pattern is as follows: [], one of the amino acids; x, any of the amino acids; (), number of repeats; (x, y) from x number of repeats to y number of repeats; {}, none of the listed amino acids.

Table 1 shows that phosphorylation consensus patterns C1, C3, C4 and C5 in the test set appear more outside than inside, which indicates that consensus patterns C1, C3, C4 and C5 are poor pattern

[2] 5 out of 62 sequences could not be trained by the reconstructed TMHMM.

predictors (marked in bold). The values (in, out and helix) for each consensus pattern do not add up to 1 because they were normalized by the total length of sequences' being inside, outside and in the helix of the membrane respectively. Though consensus patterns C2 and C10 tend to follow the assumptions, they are still not specific enough for TM protein topology prediction. Next, we examined NetOGlyc and NetPhos.

3.5.1.1.2 NetOGlyc 2.0/NetPhos 2.0

NetOGlyc [36] is a tool for the prediction of type O-glycosylation sites in mammalian proteins and NetPhos [37] is a tool for the prediction of serine (S), threonine (T) and tyrosine (Y) phosphorylation sites in eukaryotic proteins. NetOGlyc 2.0 was used with a potential greater than 0.9 on the subset of 64 mammalian proteins and NetPhos 2.0 was used with a score greater than 0.9 on the subset of 76 eukaryotic proteins from the 160 data set.

However, incorportation of NetOGlyc 2.0 and NetPhos 2.0 prediction results did not generate desirable results either. There are many false positives. Study on the loop amino acid composition of TM protein indicates that the high content of threonine on the extracellular side is not caused by glycosylation only [14]. This might be one of the reasons NetOGlyc 2.0 failed. Instead, we chose to use PROSITE signatures and domains to augment TMHMM.

3.5.1.1.3 Internal and external domains

From query against PROSITE release 17.4, we found 3 internal (including 2 phosphorylation signatures) and 1 external domains in our training data set. The details are listed in Table 2.

Table 2: Signatures obtained non-computationally from PROSITE in the 160 data set.

signature	specificity[1]	sensitivity[2]
A4_INTRA (Amyloidogenic glycoprotein intracellular domain signature) (assumed internal) G-Y-E-N-P-T-Y-[KR]	100.00%	100.00%
A4_EXTRA (Amyloidogenic glycoprotein extracellular domain signature) (assumed external) G-[VT]-E-[FY]-V-C-C-P	100.00%	100.00%
PTS_EIIB_CYS PTS EIIB domains cysteine phosphorylation site signature (assumed internal) N-[LIVMFY]-x(5)-C-x-T-R-[LIVMF]-x-[LIVMF]-x-[LIVM]-x-[DQ] [C is phosphorylated]	100.00%	96.67%
PTS_EIIA_2 PTS EIIA domains phosphorylation site signature 2 (assumed internal) [DENQ]-x(6)-[LIVMF]-[GA]-x(2)-[LIVM]-A-[LIVM]-P-H-[GAC]	100.00%	93.10%

[1] specificity: value is from PROSITE.

[2] sensitivity: value is from PROSITE.

However, incorporation of them into AHMM did not make apparent improvement on both the training and test set. Proteins in the test set do not contain any of them. We then sought computational approach to see if we could make any improvement over TMHMM.

3.5.2 Computational approach

In order to extend the set of functional domains, we use computational approach to choose specific signatures and domains that are not phosphorylation and glycosylation motifs, but are located preferentially internal or external to the membrane.

The selection was conducted as follows:

1) Run the training sequences against PROSITE database to obtain the corresponding signature(s) and/or domain(s) for each sequence with profile cut-off level L = 0 (trusted cut-off for positive matches).

2) Check for each PROSITE signature and domain contained in the training sequences to see where it resides, for example, inside (cytoplasmic), outside (exoplasmic) of the membrane or in helix and how many non-redundant sequences (incidences) correspond to it.

3) If a signature or domain appears exclusively inside or outside of the membrane at least twice, it is selected for further test.

4) Incorporate all signatures and domains selected from step 3) into Viterbi algorithm and exclude all the signatures and domains that cause an error during the prediction (with profile cut-off level L = 0, only one pattern caused an error, namely, the ATP/GTP-binding site motif A ATP_GTP_A. However, with L = –1, more patterns caused errors than with L = 0). The remaining signatures and domains are the potential predictors. They are then tested on the test sequences. The potential signature and domain predictors extracted from 157[3] sequences are shown in Table 3.

[3] There are 3 sequences in the 160 data set, which TMHMM predicted correctly whereas the reconstructed TMHMM predicted wrongly. We excluded these sequences from the training set to extract potential functional domain predictors for the fairness of comparison.

Table 3: Potential signature and domain predictors extracted from 157 sequences and tested on 62 sequences.

signature and domain	sidedness	specificity[1]	sensitivity[2]	P_H	for 157		for 62	
					better	worse	better	worse
?NEUROTR_ION_CHANNEL (Neurotransmitter-gated ion-channels signature)	external	100.00%	99.43%	0.6	1	0	appears: 0 time	
					appears: 12 times (GRA1_HUMAN)			
PROTEIN_KINASE_ATP (Protein kinases ATP-binding region signature)	internal	96.25%	84.94%	0.6	1	0	appears: 0 time	
					appears: 4 times (CEK2_CHICK)			
PROTEIN_KINASE_TYR (Tyrosine protein kinases specific active-site signature)	internal	94.79%	98.41%	0.6	3	0	appears: 0 time	
					appears: 4 times (CEK2_CHICK, EGFR_DROME, EGFR_HUMAN)			
HIS_KIN (Histidine kinase domain) [profile]	internal	100.00%	100.00%	0.6	2	0	1	0
					appears: 3 times (ENVZ_ECOLI, PHOR_ECOLI)		appears: 2 times (CPXA_ECOLI)	
PRO_RICH (Proline-rich region) [profile]	internal	*		0.6	0	0	1	0
					appears: 1 time		appears: 2 times (SCAA_RAT)	
?CONNEXINS_1 (Connexins signature 1)	external	100.00%	91.80 %	0.6	appears: 3 times		appears: 0 time	
?CONNEXINS_2 (Connexins signature 2)	external	100.00%	100.00%	0.6	appears: 3 times		appears: 0 time	
C_TYPE_LECTIN_1 (C-type lectin domain signature)	external	89.05 %	70.93 %	0.6	appears: 3 times		appears: 0 time	
SPASE_I_3 (Signal peptidases I signature 3)	external	70.59 %	94.74 %	0.6	appears: 2 times		appears: 0 time	
PROTEIN_KINASE_DOM (Protein kinase domain) [profile]	internal	99.71%	99.63%	0.6	appears: 4 times		appears: 0 time	
C_TYPE_LECTIN_2 (C-type lectin domain) [profile]	external	98.48 %	98.48 %	0.6	appears: 4 times		appears: 0 time	
?ARG_RICH (Arginine-rich region) [profile]	internal	*		0.6	appears: 2 times		appears: 0 time	
AAA (AAA-protein family signature)	internal	100.00%	96.86%	0.6	1	0	appears: 0 time	
					appears: 1 time (FTSH_ECOLI)			

* Proline-rich region can, in some cases, be ignored by a program (because it is too unspecific) (quoted from PROSITE).

[1] specificity: value is from PROSITE.

[2] sensitivity: value is from PROSITE.

P_H is the probability that the signature or domain is internal or external based on the assumption. In this experiment, to be conservative, we set it as 0.6 because we do not know exactly what the value is.

Functional domains PROTEIN_KINASE_ATP (Protein kinases ATP-binding region signature), PROTEIN_KINASE_DOM (Protein kinase domain [profile]), and PROTEIN_KINASE_TYR (Tyrosine protein kinases specific active-site signature) appear at the same time in the examined sequences. Functional domain CYTOCHROME_C is taken out to avoid false positives because its specificity is only 43.11% according to PROSITE. Signature AAA is added because it is internal according to the expert's opinion [48] and it could help in prediction of sequence FTSH_ECOLI. Domain PRO_RICH was obtained with profile cut-off level $L = -1$ (a match is potential (weak), especially if there are other matches in the sequence with the profile) and was confirmed by the expert to be internal. Sequence names marked by green are sequences predicted wrongly by TMHMM but correctly by AHMM.

Those signatures and domains without any mark in front of their names are confirmed with the expert's opinion. Those with question marks are unknown yet (needs more investigation).

3.6 Comparison between TMHMM and AHMM

AHMM incorporated with consensus patterns generates fairly poor prediction result. However, AHMM incorporated with signatures and domains gives a much better result. The AHMM here refers to AHMM incorporated with PROSITE signature and domain predictors extracted from 157 training sequences. Comparisons can be done at two levels: amino acid level and sequence level.

3.6.1 Amino acid level

It is the percentage of overlap between topologies predicted by either TMHMM or AHMM and the reference topology for sequence with functional domain predictors in labeling (“i” stands for inside of the membrane; “o” stands for outside of the membrane and “M” stands for helix).

3.6.2 Sequence level

It is the correctness on both helix number and orientation between topologies predicted by either TMHMM or AHMM and the reference topology for sequences with functional domain predictors. In detail, for each helix in the reference topology, if at least 5 amino acids in the prediction overlap with it, we believe at sequence level the helix prediction is correct. If the N-terminus orientation is also correct, then the prediction is correct.

For example, for sequence GRA1_HUMAN, the topology labeling from reference, TMHMM (predicted incorrectly) and AHMM (predicted correctly) are shown below:

Reference (4 helices and N_{out}-C_{out} topology):

```
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
ooooooooooooooooooooooooMMMMMMMMMMMMMMMMMMMMMMMMMiiiiiMMMMMMMMMMMMMMMM
MMoooooooooooooMMMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMoooooo
ooooo
```

TMHMM (wrong topology, 3 helices and N_{in}-C_{out} topology):

```
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMMMooooooooooooooooooooooooo
ooooooooooooooooMMMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMoooooo
ooooo
```

AHMM (correct topology with the aid of Neurotransmitter-gated ion-channels signature underlined):

```
ooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
ooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
ooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
ooooooooooooooooooooooooooooMMMMMMMMMMMMMMMMMMMiiiiiiMMMMMMMMMMMMMMMMMMMMoo
ooooooooooooooooMMMMMMMMMMMMMMMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMoooooo
ooooo
```

3.7 Test for the robustness of AHMM

We incorporated the potential signature and domain predictors extracted from 157 sequences into Viterbi algorithm and tested on the 62 sequences. With profile cut-off level L = 0, we found one sequence (CPXA_ECOLI) that was predicted wrongly by TMHMM but was predicted correctly by

AHMM. However, with profile cut-off level L = −1, we found two sequences (CPXA_ECOLI, SCAA_RAT) that were predicted wrongly by TMHMM but were predicted correctly by AHMM.

In order to test the robustness of the method, we re-sampled and evaluated a total of 219 sequences (the 157 training plus 62 test sequences) twenty times at both amino acid level and sequence level . That is: select 157 non-redundant random samples as training data and the rest 62 sequences as test data. Then, conduct the computational selection of signatures and domains and test them altogether on the test set. We repeated this twenty times. Only the test results are shown below (Table 4).

Table 4: Results of twenty time-re-samplings on 219 sequences. Column TMHMM2.0 is the percentage of correctly predicted amino acids by TMHMM 2.0 over sequences with potential PROSITE functional domain predictors; similarly, column TMHMM1.0 is the percentage of correctly predicted amino acids by TMHMM 1.0 and column AHMM is the percentage of correctly predicted amino acids by AHMM.

run	62 (amino acid level)			62 (sequence level)				
	TMHMM2.0	TMHMM1.0	AHMM	same	better[1]	worse[2]	ISD[3]	# of seqs
1	0.9685	0.8020	0.9783	8	1	0	13	9
2	0.7584	0.7910	0.9861	5	2	0	9	7
3	0.9072	0.9173	0.9325	6	2	0	11	8
4	0.9023	0.9257	0.9784	8	1	0	15	9
5	0.8193	0.8490	0.9869	5	2	0	11	7
6	0.8400	0.6904	0.8618	6	2	2	12	10
7	0.7942	0.7959	0.9707	9	3	0	16	12
8	0.7424	0.7764	0.9718	5	2	0	13	7
9	0.8613	0.8951	0.9717	7	1	0	11	8
10	0.6978	0.7019	0.9715	6	3	0	15	9
11	0.8520	0.6708	0.9835	4	2	0	12	6
12	0.8736	0.8719	0.9499	10	2	0	17	12
13	0.7652	0.7843	0.9648	9	3	0	18	12
14	0.8717	0.7696	0.9792	9	3	0	19	12
15	0.7823	0.6172	0.9754	5	3	0	16	8
16	0.7612	0.7907	0.9765	5	2	0	12	7
17	0.7042	0.7318	0.9888	5	3	0	11	8
18	0.9618	0.8012	0.9804	7	1	0	11	8
19	0.8610	0.8615	0.9884	4	1	0	9	5
20	0.7885	0.7870	0.9857	5	2	0	10	7
wavg[4]	0.8279	0.7922	0.9675		0.2398	0.0117		

[1] better—the sequence where TMHMM predicted wrongly but AHMM predicted correctly

[2] worse—the sequence where TMHMM predicted correctly but AHMM predicted wrongly

[3] ISD—number of signatures and domains identified

[4] wavg (weighted average) = total number of correctly predicted amino acids / total length of all sequences with functional domains.

Column "# of seqs" lists the actual number of sequences for comparison between TMHMM and AHMM at each run. The actual number of sequences for comparison depends on the number of sequences containing functional domains.

For each run, a weighted average for the percentage of overlap with reference labeling was calculated for all the sequences with functional domains. At the end, a weighted average was calculated for all the twenty runs. From Table 4, we can see that at both sequence level [4]and amino acid level[5], AHMM is better than TMHMM. On average, AHMM is better than TMHMM by more than 10% at amino acid level for sequences with functional domains. This result is also verified by a four time-5-fold cross-validation.

Worse cases occurred when the signature appeared on the different side of the membrane in the test data than it was in the training data (i.e. ATP/GTP-binding site motif A ATP_GTP_A).

Functional domains for the above experiment were obtained from PROSITE release 17.4 of May 2002 with profile cut-off level L = 0.

Statistical tests were designed to test the results for each run of resampling at amino acid level. The hypothesis is that there is no difference between AHMM and TMHMM. For each amino acid of a TM protein, it can be predicted either correctly or incorrectly with respect to the reference labeling. Thus, a sequence of amino acids can be seen as a sequence of binomial trials. Since the population of TM proteins might not be normally distributed, non-parametric tests, sign test and Wilcoxon Matched-Pairs Signed-Ranks Test were conducted at each run to compare prediction results between TMHMM and AHMM for sequences with functional domains. On the other hand, since the number

[4] For each helix in the reference topology, if at least 5 amino acids in the prediction overlap with it, we believe at sequence level the helix prediction is correct. If the N-terminus orientation is also correct, then the prediction is correct.

of amino acids at each run is large enough, t-tests for Paired Samples were also conducted. All statistical tests were run with SPSS release 6.1. 1-tail Ps and 1-tail significances of all the statistical tests between AHMM and TMHMM are less than 0.01. This indicates that if the null hypothesis is true, the chance of getting such sample difference in Table 4 is $P<0.01$. Therefore, we should reject the null hypothesis and conclude that AHMM is better than both versions of TMHMM for sequences with functional domains.

3.8 Sensitivity and Specificity of TMHMM and AHMM on helix and sidedness prediction

In addition to the above experiments, we further tested the sensitivity and specificity of TMHMM and AHMM on helix and outsidedness prediction over sequences with functional domains out of 62 from the twenty time-resampling (Table 5).

[5] the percentage of overlap with the reference topology

Table 5: Comparison of sensitivity and specificity between TMHMM and AHMM on helix and outsidedness prediction for sequences with functional domains out of 62 from each resampling. SEH is the sensitivity for helix prediction at sequence level; SPH is the specificity for helix prediction at sequence level; SEO is the sensitivity for outsidedness prediction at amino acid level and SPO is the specificity for outsidedness prediction at amino acid level.

run	SEH			SPH			SEO			SPO		
	T2.0[1]	T1.0[2]	AHMM	T2.0	T1.0	AHMM	T2.0	T1.0	AHMM	T2.0	T1.0	AHMM
1	1.0	1.0	1.0	0.9677	0.9677	1.0	0.9849	0.8396	0.9948	0.9824	0.8116	0.9847
2	1.0	0.9333	1.0	1.0	1.0	1.0	0.8821	0.9393	0.9969	0.7757	0.7838	0.9937
3	0.8929	0.9286	1.0	1.0	1.0	1.0	0.8756	0.8791	0.8967	0.9463	0.9653	0.9860
4	1.0	0.96	1.0	1.0	1.0	1.0	0.9424	0.9925	0.9903	0.8908	0.8933	0.9895
5	0.9286	0.8571	1.0	1.0	1.0	1.0	0.8498	0.9001	0.9964	0.8707	0.8753	0.9948
6	0.8718	0.8718	0.8205	0.9189	0.9189	0.9143	0.8588	0.7129	0.8009	0.8467	0.6831	0.9687
7	0.9211	0.9211	0.9737	0.9459	0.9459	0.9487	0.8583	0.8587	0.9834	0.7796	0.7792	0.9828
8	1.0	0.95	1.0	0.9524	0.95	1.0	0.9281	0.9934	0.9904	0.7067	0.7196	0.9834
9	1.0	0.9524	1.0	1.0	1.0	1.0	0.9378	0.9940	0.9914	0.8746	0.8778	0.9867
10	0.9655	0.9655	1.0	0.9655	0.9655	1.0	0.9238	0.9258	0.9901	0.6177	0.6190	0.9780
11	0.9412	1.0	1.0	0.8889	0.8947	1.0	0.9770	0.8078	0.9885	0.8070	0.6580	0.9960
12	0.9286	0.9286	1.0	1.0	1.0	1.0	0.9025	0.9027	0.9458	0.9138	0.912	0.9907
13	0.9429	0.9143	0.9714	0.9167	0.9143	0.9444	0.9389	0.9804	0.9772	0.7179	0.7252	0.9841
14	0.9231	0.9487	1.0	0.9730	0.9737	1.0	0.9143	0.8155	0.9921	0.8629	0.7629	0.9883
15	1.0	1.0	1.0	0.9130	0.9130	1.0	0.9232	0.7795	0.9931	0.7407	0.6064	0.9833
16	0.9412	0.8824	0.9412	0.8421	0.8333	0.8889	0.9307	0.9836	0.9811	0.7428	0.7520	0.9904
17	1.0	0.9375	1.0	0.9412	0.9375	1.0	0.8984	0.9459	0.9950	0.7046	0.7137	0.9950
18	0.9545	1.0	1.0	0.9545	0.9565	1.0	0.9727	0.7972	0.9904	0.9660	0.7704	0.9856
19	1.0	1.0	1.0	0.9333	0.9333	1.0	0.9938	0.9979	0.9979	0.7728	0.7727	0.9918
20	0.9286	0.9286	1.0	1.0	1.0	1.0	0.8422	0.8422	0.9988	0.8705	0.8686	0.9927
wavg	0.9502	0.9419	0.9793	0.9542	0.9538	0.9813	0.9131	0.8935	0.9737	0.8120	0.7747	0.9875

[1] T2.0—the sensitivity or specificity of TMHMM 2.0 on helix or outsidedness prediction.

[2] T1.0— the sensitivity or specificity of TMHMM 1.0 on helix or outsidedness prediction.

SEH is the number of helices predicted correctly compared with the reference helix number and SPH is the number of helices predicted correctly compared with the predicted helix number. SEO is the number of amino acids predicted correctly compared with the reference number of amino acids being outside and SPO is the number of amino acids predicted correctly compared with the predicted number of amino acids being outside. Weighted average was conducted for each run and for all twenty runs.

Table 5 illustrates that AHMM is more specific and sensitive than TMHMM on helix and sidedness prediction for sequences with PROSITE functional domains. AHMM is especially more specific and sensitive than TMHMM on sidedness prediction. At run 6, lower sensitivity and specificity on helix prediction and lower sensitivity on sidedness prediction of AHMM compared with TMHMM corresponds to the errors occurred in run 6 of Table 5. Nevertheless, the specificity on sidedness prediction of AHMM is still higher than TMHMM.

3.9 Comparison between ‘fixed helix HMM’ and AHMM

Comparison between fixed helix HMM and AHMM was conducted at sequence level. Both methods used Table 3 signature and domain predictors (except Pro-rich region and AAA-protein family signature). In Table 6, six sequences from 157 sequences were predicted correctly by AHMM but not by TMHMM and ‘fixed helix HMM’. Furthermore, there is one sequence (CPXA_ECOLI) which was predicted wrongly by TMHMM but was predicted correctly by both ‘fixed helix HMM’ and AHMM at L = 0. At L = –1, both CPXA_ECOLI and SCAA_RAT were predicted correctly by AHMM, but not by ‘fixed helix HMM’.

Table 6: Comparison between two different implementation methods on TM protein topology prediction (L = 0).

data set	fixed helix HMM		AHMM	
	better	worse	better	worse
157	0	0	6	0
62	1	0	1	0

In summary, fixed helix HMM is good in correcting sidedness errors whereas AHMM are good in correcting both sidedness and helix number errors. However, they all depend on good predictors.

Chapter 4

Discussions and Conclusion

4.1 Discussions

From our experiments (Table 4 and Table 5), we found that implementation of functional domains on top of TMHMM can improve TM protein prediction accuracy at both sequence level and amino acid levels. Furthermore, it improves both sensitivity and specificity on helix and sidedness prediction. From Table 6, we could see that AHMM outperforms the 'fixed helix HMM', since it fixes not only sidedness errors, but also helix number errors. In summary, sidedness is not decided by N-terminus alone. Sidedness and helix position are not two independent issues. Therefore, topology should be examined as a whole. Following are some discussions on the GeneomeScan formula, functional domains, proteins, the scope of AHMM and protein structure prediction techniques.

Two observations are obtained regarding the GenomeScan formula. When we used NetOGlyc 2.0 results to help to predict TM topology, we observed that the GenomeScan formula (presented in Section 3.3) was sensitive to NetOGlyc prediction errors. For example, for sequence GLP_PIG, TMHMM predicts correctly. However, after incorporation of NetOGlyc 2.0 prediction results (though 5 out of 7 are correct) into AHMM, we have the wrong topology instead of the correct topology. Here is a closer examination:

TMHMM (correct topology):

ooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooMMMMMMMMMMMMM
MMMMMMMMMMiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii—P1, P3

AHMM (with NetOGlyc 2.0 prediction results, wrong topology):

iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMM
MMMMMMMMMMooooooooooooooooooooooooooooooooooooooooooooooo—P2, P4

Let P1 be the TMHMM (Viterbi) probability for the correct topology above, P2 be the TMHMM probability for the wrong one, P3 be the AHMM (adjusted) probability for the correct topology, P4 be the AHMM probability for the wrong one.

There are 5 threonines (T) at position 1, 6, 21, 118, and 130 and 2 serines (S) at position 11 and 28. Threonines and serines were the predicted glycosylation sites and were assumed to appear external to the membrane. Here is the analysis:

For TMHMM,

P1 > P2

For AHMM, after incorporation of the prediction results of NetOGlyc 2.0,

P3 < P4

P3 = P1 * increased 5 times * decreased 2 times

P4 = P2 * decreased 5 times * increased 2 times

$P(\Phi_H)$ at position 118 dropped dramatically. This caused the probability of being outside of the membrane at position 118 to increase dramatically. Even though NetOGlyc2.0 made only two mistakes (position 118 and 130), we did not obtain the correct prediction. Predictors containing wrong information may cause wrong prediction. However, modification of the existing formula may also be needed.

Next, there is certain subjectivity in the choice of the value of the probability or weight P_H for functional domains in the GenomeScan formula. As mentioned earlier, we set $P_H = 0.6$ for all the functional domains incorporated into AHMM. We might need to incorporate a more refined and accurate number. Nevertheless, we also tried $P_H = 0.9$, which made no difference compared to 0.6. This might suggest that the functional domains in the experiment are fairly specific.

The following are a few observations upon selection of functional domains. If a pattern is not specific enough (i.e. short), incorporation of such a pattern may cause many false positives. Sequences predicted correctly by TMHMM could even be predicted wrongly. To solve this problem, we used specific patterns—signatures and domains of PROSITE instead. They are typically longer than consensus patterns of PROSITE. Matching with them is less likely to be random.

We used ps_scan, a perl program to scan PROSITE locally. There are two profile cut-off levels in ps_scan: L = 0 and L = –1. With L = 0, all hits are true positives, but false negatives may be missed. On the other hand, with L = –1, all true positives are covered, but false positives may also be included. To be conservative, we chose L = 0. With profile cut-off level L = –1, there were more potential functional domain predictors than with L = 0. However, there were more poor functional domain predictors as well. It is hard to find the optimal solution.

Another observation is that there are some amino acid-rich domains, such as PRO_RICH (proline-rich region) and ARG_RICH (arginine-rich region). In PROSITE, these domains were said to have low specificity; in our study, however, they did not cause any false positives. In fact, using PRO_RICH even helped in correcting wrong topologies predicted by TMHMM.

In addition to the analysis of functional domains in loops of TM proteins, we also examined functional domains in helix region. GLYCOPHORIN_A is the only transmembrane domain found in

219 sequences with PROSITE release 17.4 at L = 0. However, incorporation of GLYCOPHORIN_A into AHMM did not make any apparent improvement with the current data.

With regard to data sets, two observations merit further discussion. In our experiments, we put both prokaryotic and eukaryotic membrane proteins together. This is due to 1) The number of known topology membrane protein sequences is limited; 2) The training data used to extract functional domains includes both prokaryotic and eukaryotic membrane protein sequences.

In the future, prokaryotic and eukaryotic membrane proteins and organelle membrane proteins should be trained and tested separately. They all have different lipid environment, membrane height and translocation machineries which impose different constraints on membrane insertion. One notable example is TopredII. It predicts eukayotic and prokaryotic membrane proteins differently.

During the experiment, TonB protein caught our attention. According to Swiss-Prot [49], TonB protein is said to be "anchored to the cytoplasmic membrane via its n-terminal signal-like sequence, spans the periplasm" (exact quote from Swiss-Prot). However, recent studies suggest that TonB protein shuttles between the cytoplasmic membrane and outer membrane in *E. coli*. The most interesting aspect is that its N-terminal signal anchor can detach from the cytoplasmic membrane during energy transduction and becomes associated solely with outer membrane [38]. This peculiar behavior of TonB may indicate that the insertion mechanism for TonB protein is different than that of other inner membrane proteins of Gram-negative bacteria. If this hypothesis is validated, TonB should not be included in the test set.

On the other hand, a proline-rich region exists and was found in TonB protein. This proline-rich region can help to identify the TM segment of TONB_ECOLI. Proline-rich region is believed to be a cytoplasmic domain. However, in TonB protein, the proline-rich region is periplasmic. Whether this is due to the low specificity of proline-rich region or the uniqueness of TonB insertion mechanism requires more study.

We also have an important observation on AHMM. Patterns and domains studied in AHMM were from native integral membrane proteins. Thus, AHMM is not valid for predicting artificial membrane proteins. By redistributing positively charged amino acids in the loops, the topologies of artificially engineered membrane proteins are altered. Functional domains reside on one side of the membrane could end up on the different side of the membrane. One example of the artificial membrane proteins is the fusion protein LEP-LEP, which is constructed from *E.coli* inner membrane leader peptidase (LEP).

LEP has two TM segments and a N_{out}-C_{out} toplogy. The loop containing the PROSITE signature SPASE_I_3 (Signal peptidases I signature 3) of LEP is on the external side of the membrane. However, by introducing 3 lysines (K) to the 2nd loop of LEP-LEP, the mutant adopts "leave one out" topology and the loop containing signature SPASE_I_3 appears on the internal side of the membrane [39].

Upon comparing the prediction techniques between membrane protein topology prediction and soluble protein structure prediction, we have the following observations: The major difference between membrane protein topology prediction and soluble protein structure prediction is on the sidedness prediction. Membrane protein prediction must address loop sidedness whereas soluble protein prediction does not involve loop sidedness. In general, there are three different techniques for soluble protein 3D structure prediction: namely, homology modeling, protein threading and *ab initio* folding. Homology modeling is based on the idea that the structure of a protein is similar to its homologous proteins. Protein threading is based on the idea that structures are conserved among certain divergent sequences. *Ab initio* folding is based on the sequence only [43].

Currently, there are only 77[6] integral membrane proteins with 3D structures according to the White Laboratory as of Jan. 20, 2004 [47]. So far, only two simple folds were found: the helical bundle and the closed beta barrel [44]. Thus, protein threading may not be suitable for membrane protein structure prediction since there are a very limited number of 3D structures available for templates.

AHMM uses the following idea: if test proteins contain the same signature and/or domain predictors as the training proteins, then the signatures and/or domains of the test proteins will tend to be on the same side of the membrane as they are in the training proteins. In other words, the topologies of the corresponding part of the test proteins are the same as those of the training proteins. PROSITE signatures are conserved regions of proteins obtained through sequence alignments whereas domains are derived from extremely divergent sequences by PSSM. In this way, AHMM used homology modeling on signature information and protein threading on domain information for membrane protein topology prediction. Homology modeling was also used by PHDhtm during multiple sequence alignments. It is believed that homologous proteins have approximately equal secondary structures if there is 25-30% sequence similarity [26].

On the other hand, TMHMM (or AHMM) is similar to threading in the sense that it predicts membrane protein topology not necessarily from the same family of proteins where it was trained. Only if the HMM is trained from a specific family of membrane proteins and predictions are made on the other member proteins from the same family, it is homology modeling.

As far as implementation of pairwise interaction or helix-helix interaction is concerned, it might not be reliable to use protein threading method. Nevertheless, we could use 3D structure prediction tools to verify the prediction results of 2D structure prediction tools on TM segments.

[6] Includes proteins of same type from different species. For example, photosynthetic reaction centers from *R. viridis* and *R. sphaeroides* are considered unique.

Last but not all, it must be noted that the reference topology is for reference only. Even the crystal structures of TM proteins do not elucidate the exact boundaries of the protein in the lipid bilyer. This is a challenge for TM protein topology prediction.

4.2 Future work

There is still a substantial amount of sequences that are predicted wrongly. We need to further improve the prediction accuracy on TM protein topology prediction. There are three possible ways to improve.

First, change the current HMM architecture to fit better with the biological features and insertion mechanism of membrane proteins and train a new version of HMM on more TM protein sequences with known topology.

Second, use a better-developed model other than HMM to implement helix-helix interactions and other features and mechanisms characterized in the future. Helix-helix interaction is among one of the rigorous research areas in TM proteins. Helix-helix interaction is a definite phenomenon observed among helices of TM proteins. Practically, to implement this model, we have to remember what have seen before over a variable length of amino acids. Through implementation of helix-helix interaction, we may be able to better recognize the helix, which does not have strong topogenic signals and may otherwise be overlooked by current prediction methods.

Third, train and test specific models. For example, train and test eukaryotic, prokaryotic and organelle membrane proteins separately. We could also train and test models especially for ion channels and GPCP proteins. They are difficult to predict because their TM segments contain high proportion of polar residues [1].

Currently there are many biological mechanisms that still remain unknown. This has caused some major difficulties in modeling TM protein topology. For instance, do start and stop transfer events really exist? What makes them as signals? Is it simply because of hydrophobicity or are there any other factors involved?

We found more functional domains when searched against InterProScan [45]. However, none of them helps to make any apparent improvement. Besides, a fair amount of domains have not been completely annotated by InterProScan version v3.1.

Furthermore, only a fraction of sequences have PROSITE functional domain predictors. As more and more sequences with known topology are available, we would expect more useful predictors (including those which were filtered out at present) could be found in the future. We also would expect that as more and more signatures and domains are available, the prediction accuracy would be further improved with more potential predictors. With L = 0, PROSITE release 18.9 of 4-Oct-2003 was compared with release 17.4 of May 2002. We found more functional domains (i.e. IG_LIKE Ig-like domain profile) and predicted one more sequence (MYP0_HUMAN) correctly. However, these expectations still need to be further proven.

In addition, if we have more TM proteins with known topology, we would know more on the length distribution of loop and helix. For example, in TMHMM, the length for a loop before and after entering into the globular region is 10 amino acids long. This set-up has not been verified biologically.

Appendix A

Glossaries

Cleavage site: the cleavage site of a signal peptide is recognized and cleaved by the signal peptidase on the luminal side of the ER or extracellular side of the plasma membrane.

cDNA (complementary DNA): single-stranded DNA complementary to an RNA, synthesized from it by reverse transcription *in vitro*.

Compositional distance: A protein or peptide is represented as one point in amino acid composition space [40]. The distance between two proteins, j and k is calculated by $d_{jk} = [\sum_{i=1}^{20} (CA_{ij} - CA_{ik})^2]^{1/2}$ where $CA_i = A_i / SD_i$ is the normalized composition of a protein, A_i is the percentage of amino acid type-i and SD_i is the standard deviation over a large set of proteins.

Gene fusion: the use of recombinant DNA techniques in generating hybrid or chimeric polypeptides in which the tested amino acid sequence is taken from one protein and fused to another.

Integral membrane proteins: membrane proteins that extend into and sometimes completely through the membrane.

Peripheral membrane proteins: membrane proteins that lie on the surface of the membrane.

Membrane-buried helices: short hydrophobic helices, which do not span membrane lipid interior and form after stable membrane-spanning helices.

Tagging: insertion of easily identified target sites, including N-glycosylation sites, Cys residues, iodinatable sites, antibody epitopes, and proteolytic sites by site directed mutagenesis at specific positions in the polypeptide.

TM protein assembly: the process of how protein is targeted to the destined membrane and how it is inserted into the membrane.

Plasma membrane (PM) protein sorting:

Along the exocytic pathway, we are especially interested in PM protein integration. The mechanisms of protein secretion and plasma membrane protein synthesis share similarities. Exocytosis is the process in which lipid-bilayer vesicles in the cytoplasm fuse with the plasma membrane to secrete newly synthesized proteins and lipid or insert proteins into the plasma membrane. However, signals for localization of plasma membrane proteins remain unknown. Secondary sorting signals of the exocytic system are in mature polypeptides and do not seem to be related to or even contiguous with the primary signal sequence, which is responsible for the initial localization to the ER. This is quite different from most of the mitochondrial and chloroplast secondary sorting signals, and perhaps sorting in the bacterial envelope [11].

Appendix B

List of transmembrane proteins used in test set

COAB_BPFD

COAB_BPPF1

FRDC_ECOLI

FRDD_ECOLI

1B14_HUMAN

RCEH_RHOVI

RCEL_RHOVI

KCSA_STRLI

MSCL_MYCTU

COX4_PARDE

PTNC_ECOLI

PTND_ECOLI

HLYB_ECOLI

PUCC_RHOCA

BOFA_BACSU

CODB_ECOLI

LYSP_ECOLI

ARSB_ECOLI

PRRB_RHOSH

DTPT_LACLA

TRD1_ECOLI

CPXA_ECOLI

RAMP1_ECOLI

ATP6_ECOLI

TONB_ECOLI

TCR2_ECOLI

VMT2_IAUDO

HLYD_ECOLI

VNB_INBLE

MYPR_HUMAN

PMA1_NEUCR

CD7_HUMAN

GP21_RAT

ALKB_PSEOL

STE6_YEAST

FRIZ_DROME

SYB2_HUMAN

GEF_ECOLI

MSCL_ECOLI

NTG1_RAT

GSPL_PSEAE

LEP4_ERWCA

FDOI_ECOLI

FDOH_ECOLI

SCAA_RAT

FLO1_HUMAN

LCND_LACLA

CYB_RHOSH

MNTB_SYNY3

QACA_STAAU

VG1_BPFD

VRXB_LAMBD

CNG1_BOVIN

B3AT_HUMAN

ARCD_PSEAE

TOLA_ECOLI

DCTA_RHIME

GSPP_PSEAE

TAL6_MOUSE

DCRA_DESVH

CAN1_YEAST

DIVB_BACSU

Appendix C

Potential signature and domain predictors extracted from 219 sequences

signature and domain	specificity[1]	sensitivity[2]
NEUROTR_ION_CHANNEL (Neurotransmitter-gated ion-channels signature) (assumed external)	100.00%	99.43%
PROTEIN_KINASE_ATP (Protein kinases ATP-binding region signature) (assumed internal)	96.25%	84.94%
PROTEIN_KINASE_TYR (Tyrosine protein kinases specific active-site signature) (assumed internal)	94.79%	98.41%
HIS_KIN (Histidine kinase domain) [profile] (assumed internal)	100.00%	100.00%
PRO_RICH (Proline-rich region) [profile] (assumed internal)	*	
CONNEXINS_1 (Connexins signature 1) (assumed external)	100.00%	91.80 %
CONNEXINS_2 (Connexins signature 2) (assumed external)	100.00%	100.00%
C_TYPE_LECTIN_1 (C-type lectin domain signature) (assumed external)	89.05 %	70.93 %
SPASE_I_3 (Signal peptidases I signature 3) (assumed external)	70.59 %	94.74 %
PROTEIN_KINASE_DOM (Protein kinase domain) [profile] (assumed internal)	99.71%	99.63%
C_TYPE_LECTIN_2 (C-type lectin domain) [profile] (assumed external)	98.48 %	98.48 %
HLYD_FAMILY (HlyD family secretion proteins signature) (assumed external)	100.00 %	76.47 %
ARG_RICH (Arginine-rich region) [profile] (assumed internal)	*	
CYS_RICH (Cysteine-rich region) [profile] (assumed external)	*	
LYS_RICH (Lysine-rich region) [profile] (assumed external)	*	
AAA (AAA-protein family signature) (assumed internal)	100.00%	96.86%
A4_INTRA (Amyloidogenic glycoprotein intracellular domain signature) (assumed internal)	100.00%	100.00%
A4_EXTRA (Amyloidogenic glycoprotein extracellular domain signature) (assumed external)	100.00%	100.00%
PTS_EIIB_CYS (PTS EIIB domains cysteine phosphorylation site signature) (assumed internal)	100.00%	96.67%
PTS_EIIA_2 (PTS EIIA domains phosphorylation site signature 2) (assumed internal)	100.00%	93.10%

[1] specificity: value is from PROSITE.

[2] sensitivity: value is from PROSITE.

Appendix D

List of Abbreviations

AHMM:	augmented hidden Markov model
ER:	endoplamsic reticulum
GPCR:	G-protein coupled receptor
HMM:	hidden Markov model
HMMTOP:	HMM for topology prediction
MEMSAT:	Membrane protein structure and topology
PSSM:	Position specific score matrix
PHDhtm:	Profile based neural network prediction of helical transmembrane regions
PM:	Plasma membrane
SA:	Signal anchor
TM:	Transmembrane
TMHMM:	Transmembrane hidden Markov model
TOPPRED:	TOPology PREDiction program

Appendix E

Amino Acid Translation Table

Character	Translation
A	Alanine (Ala)
C	Cysteine (Cys)
D	Aspartic Acid (Asp)
E	Glutamin Acid (Glu)
F	Phenylalanine (Phe)
G	Glycine (Gly)
H	Histidine (His)
I	Isoleucine (Ile)
K	Lysine (Lys)
L	Leucine (Leu)
M	Methionine (Met)
N	Asparagine (Asn)
P	Proline (Pro)
Q	Glutamine (Gln)
R	Arginine (Arg)
S	Serine (Ser)
T	Threonine (Thr)
V	Valine (Val)
W	Tryptophan (Trp)
Y	Tyrosine (Tyr)
B	D or N (Asn or Asp)
Z	E or Q (Gln or Glu)

Bibliography

1. Moller S, Croning MDR, and Apweiler R. Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics*, 17 (7): 646-653, 2001.

2. Krogh A, Larsson B, Heijne GV and Sonnhammer ELL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol*., 305: 567-580, 2001.

3. Tusnady GE and Simon I. The HMMTOP transmembrane topology prediction server. *Bioinformatics* Application Note, 17(9): 849-850, 2001.

4. Moller S, Kriventseva EV, and Apweiler R. A collection of well characterized integral membrane proteins. *Bioinformatics* Applications Note, 16(12): 1159-1160, 2000.

5. http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/A/AnimalCells.html.

6. http://blanco.biomol.uci.edu/Membrane_Proteins_xtal.html.

7. Jones DT, Taylor WR and Thornton JM. A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry*, 33:3038-3049, 1994.

8. Martelli PL, Fariselli P, Krogh A, and Casadio R. A sequence-profile-based HMM for predicting and discriminating β barrel membrane proteins. *Bioinformatics*, 18 (Suppl.1): S46-S53, 2002.

9. http://www.gmm.gu.se/MIP-TMR/background.htm.

10. Van Geest M and Lolkema J. Membrane topology and insertion of membrane proteins: search for topogenic signals. *Microbiol Mol Biol Rev*., 64(1): 13-33, Mar. 2000.

11. Gennis RB. Biomembranes—Molecular Structure and Function. Springer-Verlag, 1989.

12. http://www.nobel.se/medicine/laureates/1999/press.html.

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

14. Nakashima H and Nishikawa K. The amino acid composition is different between the cytoplasmic and extracelluar sides in membrane proteins. *FEBS Letters*, 303(2-3): 141-146, June 1992.

15. Von Heijne G. Membrane Protein Assembly *In Vivo*. Membrane Proteins edited by Rees DC. Academic Press, 2003.

16. Lemmon MA and Engelman DM. Helix-helix interactions inside lipid bilayers. *Current Opinion in Structural Biology*, 2: 511-518, 1992.

17. http://www.mbb.yale.edu/fl/fl_d_engelman.htm.

18. Engelman DM, Steitz TA, and Goldman A. Identifying nonploar transbilayer helices in amino acid sequences of membrane proteins. *Annu. Rev. Biophys. Biophys. Chem*., 15: 321-353, 1986.

19. Claros MG and Von Heijne G. TopPred II: An improved software for membrane protein structure predictions. *CABIOS* APPLICATION NOTES, 10(6): 685-686, 1994.

20. Von Heijne G. Membrane protein structure prediction hydrophobicity analysis and the positive-inside rule. *J. Mol. Biol.*, 225: 487-494, 1992.

21. Deber CM, Wang C, Liu LP, Prior AS, Agrawal S, Muskat BL, and Cuticchia AJ. TM Finder: A prediction program for transmembrane protein segments using a combination of hydrophobicity and nonpolar phase helicity scales. *Protein Science*, 10: 212-219, 2001.

22. Juretic D, Jeroncic A, and Zucic D. Sequence analysis of membrane proteins with the web server split. CCACAA, 72 (4): 975-997, 1999.

23. Nilsson J, Persson B, Von Heijne G. Consensus predictions of membrane protein topology. *FEBS Letters,* 486: 267-269, 2000.

24. Rost B. PHD: Predicting one-dimensional protein structure by profile-based neural networks. *Methods in Enzymology,* 266: 525-539, 1996.

25. Rost B and Sander C. Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol*., 232: 584-599, 1993.

26. Rost B and Sander C. Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proc. Natl. Acad. Sci*., 90: 7558-7562, August 1993.

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

28. Durbin R, Eddy S, Krogh A and Mitchison G. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press, 1998.

29. Tusnady GE and Simon I. Principles governing amino acid composition of integral membrane proteins: Application to topology prediction. *J. Mol. Biol*., 283: 489-506, 1998.

30. Tusnady GE and Simon I. Topology of membrane proteins. *J. Chem. Inf. Comput. Sci*., 41: 364-368, 2001.

31. Sonnhammer ELL, Von Heijne G, Krogh A. A hidden Markov model for predicting transmembrane helices in protein sequences. In *Proc. Sixth Int. Conf. on Intelligent Systems for Molecular Biology,* 175-182, AAAI Press, 1998.

32. Andersson H and Von Heijne G. *Sec* dependent and *sec* independent assembly of *E.coli* inner membrane proteins: the topological rules depend on chain length. *The EMBO Journal*, 12(2): 683-691, 1993.

33. Krogh A. Two methods for improving performance of an HMM and their application for gene finding. In *Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology*, 179-186, AAAI Press, Menlo Park, CA, 1997.

34. Yeh RF, Lim LP, Burge CB. Computational inference of homologous gene structures in the human genome. *Genome Research*, 11(5): 803-806. 2001.

35. Sigrist CJ, Cerutti L, Hulo N, Gattiker A, Falquet L, Pagni M, Bairoch A, Bucher P. PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief Bioinform*, 3: 265-274, 2002.

36. Hansen JE, Lund O, Tolstrup N, Gooley AA, Williams KL and Brunak S. NetOglyc: Prediction of mucin type O-glycosylation sites based on sequence context and surface accessibility. *Glycoconjugate Journal*, 15: 115-130, 1998.

37. Blom, N, Gammeltoft, S, and Brunak, S. Sequence- and structure-based prediction of eukaryotic protein phosphorylation sites. *Journal of Molecular Biology*, 294(5): 1351-1362, 1999.

38. Larsen RA, Letain TE and Postle K. *In vivo* evidence of TonB shuttling between the cytoplasmic and outer membrane in *Escherichia coli*. *Molecular Microbiology*, 49(1): 211-218, 2003.

39. Gafvelin G and Von Heijne G. Topological “frustration” in multispanning *E.coli* inner membrane proteins. *Cell*, 77: 401-412, May 6, 1994.

40. Nakashima H, Nishikawa K and OOI T. The folding type of a protein is relevant to the amino acid composition. *J. Biochem*., 99: 153-162, 1986.

41. Blobel G. Protein targeting. *BIOSCIENCE REP*, 20(5): 303-344, Oct. 2000.

42. Chiras D. Human Biology: Health, Homeostasis and the Environment, Jones & Bartlett, 3rd edition, 1999.

43. Xu JB. Protein Structure Prediction by Linear Programming. PhD thesis, University of Waterloo, August 2003.

44. Senes A, Gerstein M and Engelman DM. Statistical analysis of amino acid patterns in transmembrane helices: the GxxxG motif occurs frequently and in association with β–branched residues at neighboring positions. *J. Mol. Biol.*, 296: 921-936, 2000.

45. Interpro. The InterPro Consortium (*R.Apweiler, T.K.Attwood, A.Bairoch, A.Bateman, E.Birney, M.Biswas, P.Bucher, L.Cerutti, F.Corpet, M.D.R.Croning, R.Durbin, L.Falquet, W.Fleischmann, J.Gouzy, H.Hermjakob, N.Hulo, I.Jonassen, D.Kahn, A.Kanapin, Y.Karavidopoulou, R.Lopez, B.Marx, N.J.Mulder, T.M.Oinn, M.Pagni, F.Servant, C.J.A.Sigrist, E.M.Zdobnov). The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucl.Acids. Res.*, 29(1): 37-40, 2001.

46. Rabiner, LR. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2): 257-286, Feb. 1989.

47. http://blanco.biomol.uci.edu/Membrane_Proteins_xtal.html

48. Personal Communications with Dr. Michael Dominguez, Cell Biologist, Caprion.

49. Boeckmann B, Bairoch A, Apweiler R, Blatter M-C, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S and Schneider M: The Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003. *Nucl.Acids. Res.*, 31: 365-370, 2003.

50. Kyte J and Doolittle RF. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, 157: 105-132, 1982.

51. Popot JL and Engelman DM. Membrane protein folding and oligomerization: the two-stage model. *Biochemistry*, 29: 17, May 1, 1990.