Chapter 4

Predicting Beta Barrel Transmembrane Proteins Using HMMs

Georgios N. Tsaousis, Stavros J. Hamodrakas, and Pantelis G. Bagos

Abstract

Transmembrane beta-barrels (TMBBs) constitute an important structural class of membrane proteins located in the outer membrane of gram-negative bacteria, and in the outer membrane of chloroplasts and mitochondria. They are involved in a wide variety of cellular functions and the prediction of their transmembrane topology, as well as their discrimination in newly sequenced genomes is of great importance as they are promising targets for antimicrobial drugs and vaccines. Several methods have been applied for the prediction of the transmembrane segments and the topology of beta barrel transmembrane proteins utilizing different algorithmic techniques. Hidden Markov Models (HMMs) have been efficiently used in the development of several computational methods used for this task. In this chapter we give a brief review of different available prediction methods for beta barrel transmembrane proteins pointing out sequence and structural features that should be incorporated in a prediction method. We then describe the procedure of the design and development of a Hidden Markov Model capable of predicting the transmembrane beta strands of TMBBs and discriminating them from globular proteins.

Keywords Hidden Markov model, Algorithms, Prediction, Membrane, Transmembrane, Beta barrel, Protein

1 Introduction

Transmembrane beta-barrels (TMBBs) constitute one of the two major structural classes of transmembrane proteins. They are located in the outer membrane of gram-negative bacteria, and in the outer membrane of chloroplasts and mitochondria. Their membrane-spanning segments are formed by antiparallel beta strands, creating a channel in the form of a barrel that spans the outer membrane [1]. The TMBBs perform a wide variety of functions such as active ion transport, passive nutrient uptake, membrane anchoring, adhesion, and catalytic activity [2–4]. Interestingly, a large number of pathogens belong to the gram-negative bacteria class, and the virulence activity in a lot of cases has been proven to depend on specific outer membrane proteins. Thus,

David R. Westhead and M.S. Vijayabaskar (eds.), *Hidden Markov Models: Methods and Protocols*, Methods in Molecular Biology, vol. 1552, DOI 10.1007/978-1-4939-6753-7_4, © Springer Science+Business Media LLC 2017

besides the obvious theoretical interest concerning the research on protein structure and function in general, this is an additional reason for attracting an increased medical interest.

Recent years have seen the development of several methods for predicting the transmembrane strands of outer membrane proteins and/or identifying these proteins in completely sequenced genomes. There is a large variation on the algorithmic techniques used for this purpose as they vary from hydrophobicity analysis [5–8], pattern recognition [9], statistical analysis using special empirical rules of amino-acid propensities and prior knowledge of the structural nature of the proteins [10], to other more refined methods including Neural Networks (NNs) [11–13], Hidden Markov Models [14–16], Support Vector Machines (SVMs) [17], Nearest Neighbors [18, 19], quadratic discriminant analysis [20], Radial Basis Function (RBF) Networks [21], Grammatical-Restrained Hidden Conditional Random Fields (GRHCRFs) [22], hybrid techniques [23, 24], and consensus approaches [25]. From a historical perspective, the prediction algorithms are divided in three categories; approaches based in hydrophobicity analysis, approaches that use statistical properties of the residues found in beta barrels and more advanced machine learning approaches. Moreover, there are two major classes of prediction methods, the methods aiming at predicting the location of the transmembrane beta strands and the methods aiming at discriminating TMBBs from other classes of proteins such as globular and alpha-helical membrane ones.

2 Structural Features of Beta Barrel Transmembrane Proteins

Transmembrane protein topology prediction has been pursued for many years in bioinformatics, mostly focusing on alpha helical membrane proteins. Alpha helical transmembrane segments are more easily predicted by computational methods, due to the easily detectable pattern of highly hydrophobic consecutive residues, and the application of simple rules as the "positive-inside rule" [26]. Another reason is the relative abundance of alpha helical membrane proteins compared to that of the beta barrel transmembrane proteins. Currently, the number of structures of beta barrel transmembrane proteins known at atomic resolution raises rapidly, due to improvements in the cloning and crystallization techniques [27], but still their numbers are significantly lower compared to their alpha-helical counterparts. Therefore, the investigation of the amino acid sequence of beta barrel transmembrane proteins and detailed examination of the available three-dimensional structures (Fig. 1) can provide distinct characteristics and general rules that may be used for the construction of a predictive method [1, 31]. Below we describe some simple rules and observations that should be considered for the design of an HMM-based method for the prediction of beta-barrel transmembrane proteins [1, 31, 32]:

- The transmembrane strands are mainly amphipathic, as they exhibit alternating patterns of hydrophobic–hydrophilic residues. The hydrophobic residues interact with the nonpolar chains of membrane lipids whereas the polar residues face toward the interior of the barrel and interact with the aqueous environment of the formed pore.
- The aromatic residues have higher abundance in the interfaces with the polar heads of the lipids. This way an "aromatic belt" is formed at the lipid bilayer interface around the perimeter of the barrel (Fig. 1).
- Both the N- and C-terminal regions are located in the periplasmic space, restricting the strand number to even values. The known three-dimensional structures contain 8–24 TM strands and in some cases, the N- and C-terminal tails of the protein may be formed by more than 100 residue-long stretches.
- The segments that connect the transmembrane strands (loops) and are located in the periplasmic space ("inside") are generally shorter than those of the extracellular space ("outside").

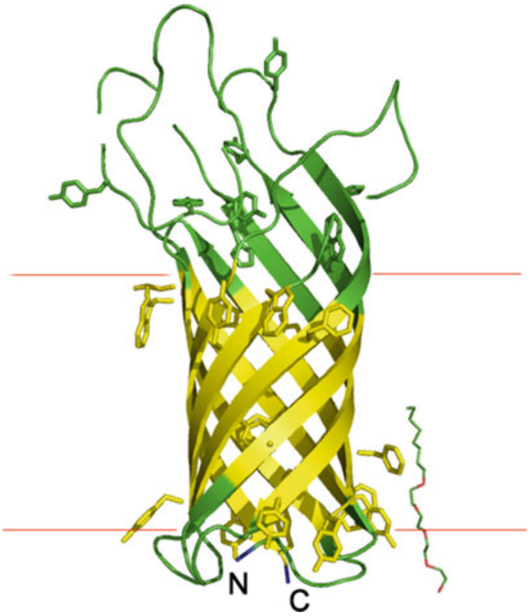

Fig. 1 A ribbon diagram of the structure of Outer membrane protein A (ompA) from *Escherichia coli* (PDB ID: 1BXW [28]). Aromatic side chains are represented as rods to illustrate the aromatic belts. The *horizontal lines* indicate the approximate position of the lipid bilayer boundaries using the annotation of PDBTM [29] for the location of the transmembrane strands. The diagram was drawn using the PyMol molecular graphics package [30]

Specifically, as observed in most three-dimensional structures, loops of the periplasmic space have approximately a maximum length of 12 amino acid residues, while extracellular loops with more than 30 residues have been observed.

- The length of the transmembrane strands varies according to the inclination angle of the strand with respect to the barrel axis (or the lipid bilayer), and ranges between 6 and 22 residues. However, in some cases, only a small portion of the strand is embedded in the lipid bilayer, while the rest of it protrudes away from the membrane towards the extracellular space, forming flexible hairpins.
- Beta barrel transmembrane proteins show high sequence variability compared to globular proteins that adopt the structure of a beta barrel. Extracellular loops exhibit the highest sequence variability in beta barrel transmembrane proteins and often function as antigenic epitopes.
- All beta strands are antiparallel and locally connected to their neighbors through a network of hydrogen bonds which stabilizes the structure of the barrel.

Below, we will provide a short review of most available methods for the prediction of beta barrel transmembrane proteins in order to describe their special characteristics in their sequence and structure that can be exploited for the design and the development of a predictive algorithm.

3 Prediction of Beta Barrel Transmembrane Proteins

3.1 Prediction Methods Based on Hydrophobicity Analysis

The alternating patterns of hydrophobic–hydrophilic residues in transmembrane strands were initially used for the prediction of transmembrane beta barrels. Vogel and Jahnig [5] calculated the average amphipathicity of each residue using a sliding window across the amino acid sequence. Jeanteur and colleagues [6], combined amphipathicity with sequence alignments of members of the protein family of porins to determine the beta strands in the porin barrel. Their approach was extended to include specific locations of aromatic residues [8] and differences between the amphipathicity of each residue and the average hydrophobicity [7]. This way, the aromatic belt of the transmembrane stands was more accurately identified. During the same period, Gromiha and Ponnuswamy [33], derived the concept of the surrounding hydrophobicity which is independent of the amphipathic features of beta strands. Years later, the Beta Barrel Finder (BBF) method [34] combined secondary structure prediction, hydrophobicity, amphipathicity, and signal peptide prediction, in order to predict transmembrane beta barrels in bacterial genomes. Based on the analysis of known three-dimensional structures they proposed that transmembrane segments could be identified as protein regions, which are predicted to form beta strands and are characterized by peaks of both hydrophobicity and amphipathicity. However, methods based on hydrophobicity analysis and/or secondary structure predictions have inherent limitations as beta sheet forming propensities and hydrophobicity scales are different between globular and transmembrane proteins [35]. Such issues should be taken into consideration when developing and using empirical predictive methods.

3.2 Statistical Approaches

Sequence features other than hydrophobicity profiles were incorporated in the prediction of transmembrane beta strands with the use of statistical approaches. Gromiha and colleagues [10], developed a set of conformational parameters for membrane spanning beta strands and described special empirical rules using amino acid propensities and prior knowledge of the structural nature of the proteins. Neuwald and colleagues [9], employed a Gibbs sampling algorithm to detect motif-encoding regions and similar repetitive motifs that characterize bacterial porins. Other approaches used multiple structural alignments [36], in order to describe certain patterns in transmembrane beta barrels that could be used for their discrimination. Wimley developed a scale-based statistical method [32] by aligning the structures with respect to the lipid bilayer and calculating amino acid propensities for residues to belong to a transmembrane beta strand or a non-transmembrane region. The original algorithm was later modified (the "Freeman--Wimley algorithm") to improve the genome scale discrimination of beta barrel transmembrane proteins [37]. On the other hand, Liu and colleagues [38] calculated differences in the amino acid composition between the two classes beta barrel membrane proteins and globular proteins for discrimination purposes. The BOMP method [39] combined regular expression patterns (in particular, a C-terminal pattern characteristic of most TMBBs) with a post processing step to filter false positives based on the overall amino acid composition, whereas the method of Wimley [32] used an additional step for the filtering of false positive predictions based on the overall amino acid composition of the protein. As the number of available three-dimensional structures continued to grow, it became obvious that the problem of the prediction of beta barrel transmembrane proteins is more complicated from the simple identification of the alternating patterns of hydrophobic–hydrophilic residues.

3.3 Machine Learning Methods

Machine Learning methods are in general more capable of capturing the nonlinear correlations of amino acids in protein sequences and perform better than statistical analyses and heuristic methods. Furthermore, the mathematical foundation of these methods is sounder and elegant predictors can be developed. Some of these methods are based solely on the amino acid sequence and others use evolutionary information derived from multiple alignments.

The first application of Machine Learning techniques for the prediction of the topology of beta barrel transmembrane proteins used a Feed-Forward Neural Network [11], to predict the relative position of C-alpha atoms with respect to the lipid bilayer. Jacoboni and colleagues [13] described the use of a similar Neural Network (B2TMPRED) which introduced the use of evolutionary information in the form of multiple sequence alignments generated by PSI-BLAST [40]. Moreover, the method included an algorithm for model optimization, based on dynamic programming in order to define more efficiently the ends of transmembrane segments and filter inconsistent predictions. A Neural Network with similar architecture, using only single sequence information as input, was presented with the TMBETA-NET method [12], which used a set of empirical rules to remove spurious predictions (e.g., strands with 3–4 residues and so on). Later, the TBBPred method [24] combined Neural Networks and Support Vector Machines (SVMs) for the prediction of transmembrane regions of beta-barrel proteins using a similar NN with B2TMPRED.

The first method based on Hidden Markov Models for the prediction of the topology of beta barrel transmembrane proteins was the HMMB2TMR method [16]. The method was trained on a dataset of 12 nonredundant outer membrane proteins, using as input multiple sequence alignments generated by PSI-BLAST. It used a HMM with different states to describe the alternating patterns of hydrophobic–hydrophilic residues in transmembrane strands, the aromatic belt located at the lipid bilayer interface and the different periplasmic and extracellular loops. The ProfTMB method [15], also included evolutionary information in the form of multiple alignments and model training and scoring procedures similar to HMMB2TMR. However, the method introduced a novel HMM architecture, different structure-based labeling, a new definition of beta-hairpin motifs, explicit state modeling of transmembrane strands, and a log-odds whole-protein discrimination score. The PRED-TMBB [41] method was the first publicly available through a web-server method for predicting the topology of TMBBs using HMMs, introducing novel training and decoding procedures. A consensus prediction method has also been described (ConBBPRED) [25], that combines the results of several methods and optimizes the predicted topology with a dynamic programming algorithm. Recently, BOCTOPUS [23] employed a hybrid method based on SVMs and HMMs and combined local per-residue predictions with global preferences. The method also incorporated a discrimination filter for more reliable identification of beta barrel transmembrane proteins.

Other methods such as TMBpro [42] and transFold [43] also predict potential interactions between the residues of beta strands that form the barrel using HMMs and multi-tape S-attribute grammars respectively. TMB-Hunt [18] employs a modified k-nearest neighbor (k-NN) algorithms to classify protein sequences as transmembrane beta-barrel. The HHomp method [44] uses a database of profile Hidden Markov Models (pHMMs) to perform sensitive sequence similarity searches through profile-profile alignments, based on the observation that almost all beta barrel outer membrane proteins have a common ancestry.

Below we describe the basic concepts for the design of a Hidden Markov Model capable of predicting the topology of beta barrel transmembrane proteins and discriminating them from other classes of proteins.

4 A Hidden Markov Model for Beta Barrel Transmembrane Proteins

4.1 Basic Concepts of Hidden Markov Models

Hidden Markov Models have been extensively used for pattern recognition problems, with the most known example found in the speech recognition methodology [45]. Hidden Markov Models have been used in bioinformatics during the last few years for generating probabilistic profiles for protein families [46], the prediction of transmembrane helices in proteins [47], the prediction of signal peptides and their cleavage sites [48], the prediction of genes [49] and for the prediction of transmembrane beta strands [14, 16, 38].

The Hidden Markov Model is a probabilistic model, which consists of several states, connected by means of the transition probabilities, thus forming a Markov process. If we consider an amino acid sequence of a protein with length L, denoted by:

$$\mathrm{x} = x_1, x_2, \ldots, x_{L-1}, x_L$$

with a labeling (in this case corresponding to transmembrane, intracellular, and extracellular regions):

$$\mathrm{y} = y_1, y_2, \cdots, y_{L-1}, y_L$$

then the transition probability for jumping from a state k to a state l is defined as:

$$\alpha_{kl} = P\left(\pi_i = l \middle| \pi_{i-1} = k\right)$$

Where π is the “path” in the particular position of the amino acid sequence (i.e., the sequence of states, as opposed to the sequence of symbols). Each state k is associated with a distribution of emission probabilities, meaning the probabilities that any particular symbol could be emitted by the current state. Assuming an alphabet Σ, consisting of the 20 amino acids, the probability that a particular aminoacid b is emitted from state k is defined as:

$$e_k(b) = P\left(x_i = b \middle| \pi_i = k\right)$$

The term "hidden" is justified by the fact that when one observes the emitted symbols he cannot see the underlying states, thus the true state process is hidden from the observer. The total probability of the observation sequence given the model, $P(\mathbf{x}|\theta)$, is computed using the efficient forward algorithm [45], whereas the joint probability of the sequence and the labeling denoted by $P(\mathbf{x},\mathbf{y}|\theta)$, is computed by its trivial modification proposed by Krogh [50].

4.2 The Hidden Markov Model Architecture

The architecture of the HMM is chosen so that it could fit as much as possible to the limitations imposed by the known structures. HMM-B2TMR uses a HMM with different types of states to describe the beta strand transmembrane core (two types), the beta strand cap on either side of the membrane, the inner loops, the outer loops, and the globular domain in the middle of each loop [16]. On the other hand, the ProfTMB method [15] uses different states to model explicitly different types of periplasmic loops (turns, hairpins etc). PRED-TMBB, used a Hidden Markov Model of similar architecture, which is described in Fig. 2. The model is cyclic, consisting of 61 states divided in three "sub-models" corresponding to the three desired labels to predict the TM (transmembrane) strand sub-model and the inner (periplasmic) and outer (extracellular) loops sub-models respectively [14].

The TM strand model incorporates states to model the special architecture of the transmembrane strands. Thus, there are states that correspond to the core of the strand and the aromatic belt located at the lipid bilayer interface. Furthermore, other states correspond to the amino acid residues facing the bilayer (the external side of the barrel) and the residues facing the barrel

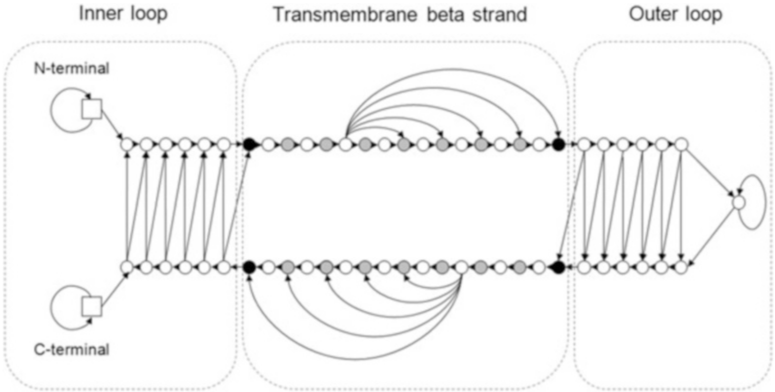

Fig. 2 A schematic representation of the model's architecture. The model consists of three sub-models corresponding to the three labels which shown separately. In the transmembrane sub-model different colors correspond to the tied states. *Black circles* correspond to the aromatic belt, gray to exterior of the strand's core, and white to the interior side. In the inner and outer loop sub-models, the states forming the ladder are tied together respectively, whereas the N-terminal tail is tied with the C-terminal and the globular outer loop state is not tied with another state. The allowed transitions are shown with *arrows*

interior. All states are connected with the appropriate transition probabilities in order to be consistent with the known structures (i.e., to ensure appropriate length distributions and to model the alternating pattern of hydrophobic–non-hydrophobic residues, corresponding to the external-internal residues of the barrel). The minimum allowed length for a transmembrane strand is seven residues, whereas the maximum is 17. However, more variable lengths can and have been used in other published methods such as HMM-B2TMR and ProfTMB.

The inner and outer loops are modeled with a "ladder" architecture of 12 states, whereas at the top of the outer loop there is a self transitioning state corresponding to residues too distant from the membrane; these cannot be modeled as loops, hence that state is named "globular." The "inner" loop sub-model has no corresponding "globular" state, reflecting the fact that inner loops are significantly shorter than the outer ones, since none of the known structures at that time, possessed an inner loop longer than 12 residues. In the future, a globular state can be added so as to cover special cases where longer loops have been observed. In order to capture the fact that all known structures are having their N-terminal region towards the periplasmic space (the "inside" with respect to the outer membrane) we allow the begin state of the model to be followed only by states belonging to the inner loop or to TM strands directing to the external side of the outer membrane. Additionally an end state can be included to fix the location of the C-terminal region at the same side. Finally, we allow a self-transitioning absorbing state to follow the inner loop states, in order to correctly model sequences that have a long C-terminus falling in the periplasmic space. States expected to have the same emission probabilities are tied together. However, different emission probabilities can be used [15] to model the strands with direction from the periplasmic to the extracellular space and strands with the opposite direction. This way the asymmetric amino acid compositions of transmembrane beta-strands [51, 52] are incorporated in the model. Finally, the emission probabilities of some of the states belonging to the extracellular loops can be modified to include the "positive outside rule" which describes higher preference of charged residues in the extracellular loops [53].

4.3 Creation of Training and Testing Datasets

For the creation of a training dataset for the HMM, a set of beta barrel transmembrane proteins with known three dimensional structure and topology must be compiled. In the original publication of PRED-TMBB [14], protein structures from the Protein Data Bank (PDB) [54] were collected after consideration of the SCOP classification [55] and in particular using the fold "Transmembrane beta-barrels." However, with the creation of specialized databases for TMBBs, such a dataset can now be compiled using information deposited in PDBTM [29], OPM [56] or TOPDB

[57] where structural and experimental data are combined. For variants of the same protein, only the structure solved at the highest resolution is collected, and multiple identical chains are removed, keeping only one chain for each structure. The sequences of the remaining structures are submitted to a redundancy check using BLAST [40], removing chains with a sequence identity above some threshold (typically 30 %).

The total number of freely estimated parameters in the model in Fig. 2 is 175. These numbers are adequate for training a prediction method using some dozens of proteins (i.e., using thousands of amino acids as observations) and in any case are significantly lower compared to the number of freely estimated parameters (weights) needed by a NN method. Most available HMM-based methods for TMBBs have used approximately 8 to 36 proteins in their training sets. Larger sets are not expected to increase the prediction performance for a given algorithmic technique. It has been shown [58] that the relationship between the sizes of the training set with the performance of the prediction algorithms is nonlinear and reaches an upper limit, even in the case of transmembrane beta barrel prediction. Following this rationale and due to the fact that the number of available three dimensional structures for transmembrane beta barrels continues to rise, a more refined training set can be compiled by collecting a structural representative from each known family of transmembrane beta barrels [59]. A complete and comprehensive collection and classification of protein families for TMBBs can be obtained from databases as Pfam [60] or OMPdb [4] where each protein family is described using pHMMs.

It is important to point out that even in structures known at atomic resolution, the exact boundaries of the transmembrane strands are not obvious, and in some situations the PDB annotations for the strands are clearly extending far beyond the membrane. Since the primary objective is to predict the TM segments of the strands rather than the entire beta strands, the model ideally must be trained to identify these particular segments. In most of the earlier works [13, 15, 16, 61], the PDB annotations for the whole strand were used but this may cause problems in the performance of the predictor. Alternatively, a manual approach described in [14] can be applied, where the labels for the TM segments are set manually, by precisely locating the aromatic belts of the barrel [31] and the residues facing the barrel interior and exterior, after inspection of the three-dimensional structures of the proteins in the training set, using molecular graphics software (Fig. 1). It is well known that discriminative training algorithms (see below) are very sensitive to data mislabeling, thus this approach was at least in part responsible for the increased performance of PRED-TMBB. More recent works however, rely on the TM assignments deposited in public databases such as PDBTM [29], which are generated by algorithms that identify the precise location of the TM part of the barrel, and thus take these considerations into account. Finally, we have to note that in [47] an automated method for relabeling the data was proposed, that consists of removing the labels at the boundaries of the TM strands, and then performing a constrained prediction that reassigns the boundaries.. This method has been used previously only in the case of alpha-helical TM proteins, and we expect that in the case of TMBBs, it will also perform well.

For the evaluation of the prediction performance of the HMM a self-consistency test and a cross-validation procedure is commonly performed. In the cross-validation procedure the training set is divided in equal subsets (folds). The model is trained with one subset of proteins and the prediction performance is evaluated against the remaining subsets. This way, prediction results for each protein in the training dataset are derived from models where the protein was not included in the training process. For small training sets measures of accuracy can be obtained using a jackknife test (leave one out cross-validation test). In addition, several test sets are compiled for the evaluation of the performance of the predictive model. An independent test set of beta barrel transmembrane proteins having experimentally verified topologies and no sequence similarity with the proteins used in the training set can be used to evaluate the topology prediction performance.

To assess the accuracy of the predictions, several measures can be used. For the transmembrane strand predictions the number of correctly predicted strands (True Positives, TP), the number of missed strands (False Negatives, FN), and the number of the over-predicted strands (False Positives, FP) are calculated. However, the strands' prediction accuracy is more efficiently measured by the segments overlap measure (SOV)Segments Overlap Measure (SOV) [62]. As measures of the accuracy per residue, the total fraction of the correctly predicted residues (Q_β or Q_3), in a two-state model (transmembrane versus non-transmembrane) or a three-state model (transmembrane versus inner loop versus outer loop), and the Matthews Correlation Coefficient (C_β) can be used [63]. In addition, the number of the correctly predicted transmembrane segments and topologies (i.e., when both strands' localization and orientation of the loops are correctly predicted) can be calculated to evaluate the performance per protein.

The discriminative power of the model can also be evaluated against a nonredundant set of globular proteins with known three-dimensional structure similar to the one used in [37] and against a negative set of alpha helical transmembrane proteins as used in [14].

4.4 Training and Decoding Algorithms

Traditionally, the parameters of a Hidden Markov Model are optimized according to the Maximum Likelihood criterion [45],

$$\hat{\theta}^{\mathrm{ML}} = \underset{\theta}{\operatorname{argmax}}\, P\left(\mathrm{x}\middle|\theta\right)$$

A widely used algorithm for this task is the efficient Baum–Welch algorithm (also known as Forward–Backward) [45, 64], which is a special case of the Expectation-Maximization (EM) algorithm, proposed for Maximum Likelihood (ML) estimation for incomplete data [65] and has been widely used for the training of several HMM methods for TMBBs [15, 16]. The algorithm, updates iteratively the model parameters (emission and transition probabilities), with the use of their expectations, computed with the use of the Forward and Backward algorithms. Convergence to at least a local maximum of the likelihood is guaranteed. The main disadvantage of ML training is that it is not discriminative. Alternatively, the Conditional Maximum Likelihood (CML) training for labeled data can be used, as proposed by Krogh [66]. Although CML is computationally more intensive than the ML approach, the predictive ability is better when data with good quality of labeling are used. The Conditional Maximum Likelihood criterion is:

$$\hat{\theta}^{\mathrm{CML}} = \underset{\theta}{\operatorname{argmax}}\, P(\mathrm{y}|\mathrm{x},\theta) = \underset{\theta}{\operatorname{argmax}}\, \frac{P(\mathrm{x},\mathrm{y}|\theta)}{P(\mathrm{x}|\theta)}$$

This kind of training, often referred to as discriminating training, seeks to maximize the probability of the correct prediction, i.e., the probability of the labeling **y** for a given sequence **x** and a model θ. Another major problem with CML is that the Baum–Welch algorithm cannot be applied and variants of the gradient descent method are needed. These methods require parameter fine-tuning (for the so-called "learning rate") and in many cases do not provide stable convergence. However, a variant that uses individual learning rates that are adapted during the process, has been presented, that offers significant advantages [67]. The CML training with the above-mentioned algorithms has been efficiently used in [14]. The parameters of the model (transition and emission probabilities) are updated simultaneously, using the gradients of the likelihood function as described in [68], and the training process terminates when the likelihood does not increase beyond a prespecified threshold. In addition, a genetic algorithm (GA) has been used [69] to train the model, demonstrating better results compared to Baum–-Welch. To reduce the number of the free parameters of the model, and thus improve the generalization capability, states expecting to have the same emission probabilities, can be tied together (Fig. 2). Furthermore, to avoid overfitting, the iterations started from emission probabilities corresponding to the initial amino acid frequencies observed in the known protein structures and small pseudocounts can be added in each step.

The decoding of an HMM can performed using the standard Viterbi algorithm [45] or alternatively the N-best algorithm [70], as formulated in [66]. This algorithm is a heuristic that attempts to find the most probable labeling of a given sequence, as opposed to the well-known Viterbi algorithm [45], which guarantees to find the most probable path of states. Since there are several states contributing to the same labeling of a given sequence, the N-best algorithm will always produce a labeling with a probability at least as high as that computed by the Viterbi algorithm, in other words it always returns equal if not better results. Its main drawback is the memory requirements and computational complexity, resulting in a slowdown of the decoding process. The original PRED-TMBB method, offered three choices: the standard Viterbi algorithm, the N-best algorith and a variant of the posterior decoder coupled with a post-processing step using dynamic programming. This algorithm has been used by Jacoboni and coworkers [13] and later presented as a general purpose algorithm [61]. However, novel decoding algorithms have been presented later, that combine the advantages of both Viterbi and posterior decoding, in a more mathematical sound way. These are the optimal accuracy posterior decoder (OAPD) [71] and the Posterior-Viterbi algorithm [72], already deployed in the topology prediction of transmembrane proteins. From our experience, these algorithms perform better compared to Viterbi, N-best, and posterior decoding, and should be preferred in future applications.

For the purpose of discrimination, the information included in the prediction of the putative transmembrane segments is not sufficient, since a prediction for a transmembrane strand could occur even in globular proteins. Thus, there is need for a global score reflecting the overall fit of the query sequence to the model. This can be derived from the negative log-likelihood of the sequence given the model, as computed by the Forward algorithm. For models trained with the ML criterion, the score is usually normalized by dividing with the likelihood of a null model, that is, a model of independence with amino acid frequencies derived from Uniprot. This is usually named “log-odds score” and is given by:

$$S(\mathrm{x}|\theta) = \frac{\log P(\mathrm{x}|\theta)}{\log P(\mathrm{x}|\mathrm{null})}$$

The log-odds score however, cannot be used for models trained with the CML criterion. In this case the likelihood is usually normalized for the length of the sequence. Thus, the statistical score used for discrimination by the PRED-TMBB method is:

$$S(\mathrm{x}|\theta) = \frac{\log P(\mathrm{x}|\theta)}{L}$$

where L is the length of the sequence. The proportion of correctly classified proteins as a function of the discrimination score used as the threshold should then be studied in order to define the optimal threshold as the value that maximizes that function. Proteins with score values below the threshold should be declared as beta-barrel transmembrane proteins. In general, the discrimination performance of such scores range at the vicinity of 90 %, and thus we expect in future applications that these scores be combined with other relative metrics (such as the number of predicted strands) in order to increase the performance.

5 Further Considerations and Improvements

As we already discussed in the respective section, improvements in prediction performance can be achieved by designing a more plausible model architecture. Currently, the potential improvements in this respect may include different emission probabilities to model the strands with direction from the periplasmic to the extracellular space and strands with the opposite direction; this way the asymmetric amino acid compositions of transmembrane beta-strands [51, 52] can be incorporated in the model. Similarly, the emission probabilities of some of the states belonging to the extracellular loops can be modified to include the "positive outside rule" which describes higher preference of charged residues in the extracellular loops [53]. Finally, the whole structure of the model may be optimized. Methods that learn the structure of the HMM using genetic algorithms, have been proposed [73, 74], and applications concerning the prediction of TMBBs may needed.

An extensive comparison and evaluation of methods for predicting the topology of beta barrel transmembrane proteins indicated that HMM-based methods outperform methods based on other types of machine learning techniques such as NNs and SVMs [25]. The regular grammar of the HMMs can capture more effectively the temporal variability of the protein sequence and map successfully the proteins modular nature to a mathematical sound model. HMM-based methods are not influenced significantly whether full-length sequences or just the beta barrel domains are submitted as input for prediction. Interestingly, the NN- and SVM-based methods, often falsely predict the signal peptide sequences as transmembrane strands in the precursors whereas HMMs do not. However, this can be solved by using a specialized signal peptide prediction method such as SignalP [75] or with the inclusion of an additional submodel in the HMM as described in [69]. Moreover, NN methods are more capable of capturing long-range correlations along the sequence. This results to the correct identification of an isolated strand, but since the beta barrel proteins follow strict structural rules, the modular nature of the barrels is captured more effectively by HMMs. NNs may often falsely predict isolated transmembrane strands in non-barrel domains or predict strands with a non-plausible number of residues or even barrels with an odd number of strands. A potential improvement in future applications can be reached by exploiting hybrid methods that combine the advantages of HMMs and those of the NNs. BOCTOPUS used such a method since it combined an SVM with a HMM-like model. However, this grammar is not a proper HMM since the transitions were not optimized, and hence there is plenty room for improvement. Such hybrid methods, are usually termed Hidden Neural Networks (HNNs) [68], but up to date, there are only a handful of applications in bioinformatics [76, 77]. Clearly, such approaches should be attractive alternatives for future applications.

Other improvements can be applied to existing HMM based prediction methods for beta barrel transmembrane proteins by utilizing algorithmic techniques and decoding algorithms that were successfully applied in the prediction of the topology of alpha helical TM proteins. Using the modifications to the standard Forward and Backward algorithms, as well as on all the above mentioned decoding algorithms for HMMs, that were extensively described in [78], prior topological information derived from experiments can be incorporated in the prediction. Subsequently, constrained predictions for TMBBs will produce more reliable topological models. Constrained predictions can also be used for the refinement of the labeling of the ends of transmembrane strands and better definition of relevant emission probabilities.

Finally, as already mentioned, some of the prediction methods use evolutionary information in the form of multiple sequence alignments [13, 15, 16, 23, 24]. The inclusion of evolutionary information in HMM based methods has been shown to substantially improve the prediction performance [16, 79]. However, most implementations used in beta barrel TM protein prediction include the construction of a sequence profile from multiple alignments that is used as an input. These methods were feasible by extending the simple HMM in a way that accepts as input a sequence profile instead of sequence. Thus, other methods such as PRED-TMBB may be difficult to adopt a similar architecture. However, a different method can be used following [71]. Briefly, given a query sequence and a multiple alignment of its homologs, predictions with the single sequence method can be obtained on each of the homologs. Then, the predicted labels can be mapped on the alignment and averaged for each position of the query sequence. This will create a "posterior label probability" (PLP) table for the query sequence that contains information from the multiple alignments. In the last step, the OAPD [71] can be applied and the final prediction is obtained. This approach, has the advantage that can be used in any single-sequence method without further modifications.

References

1. Schulz GE (2003) Transmembrane beta-barrel proteins. Adv Protein Chem 63:47–70
2. Wimley WC (2003) The versatile beta-barrel membrane protein. Curr Opin Struct Biol 13 (4):404–411
3. Bagos PG, Hamodrakas SJ (2009) Bacterial beta-barrel outer membrane proteins: a common structural theme implicated in a wide variety of functional roles. In: Daskalaki A (ed) Handbook of research on systems biology applications in medicine, pp: 182–207. doi:10.4018/978–1-60566-076-9.ch010
4. Tsirigos KD, Bagos PG, Hamodrakas SJ (2011) OMPdb: a database of {beta}-barrel outer membrane proteins from Gram-negative bacteria. Nucleic Acids Res 39(Database issue): D324–D331. doi:10.1093/nar/gkq863
5. Vogel H, Jahnig F (1986) Models for the structure of outer-membrane proteins of Escherichia coli derived from Raman spectroscopy and prediction methods. J Mol Biol 190(2):191–199, doi:0022-2836(86)90292-5 [pii]
6. Jeanteur D, Lakey JH, Pattus F (1991) The bacterial porin superfamily: sequence alignment and structure prediction. Mol Microbiol 5(9):2153–2164
7. Rauch G, Moran O (1995) Prediction of polypeptide secondary structures analysing the oscillation of the hydropathy profile. Comput Methods Programs Biomed 48(3):193–200, doi:0169260795016988 [pii]
8. Schirmer T, Cowan SW (1993) Prediction of membrane-spanning beta-strands and its application to maltoporin. Protein Sci 2 (8):1361–1363. doi:10.1002/pro.5560020820
9. Neuwald AF, Liu JS, Lawrence CE (1995) Gibbs motif sampling: detection of bacterial outer membrane protein repeats. Protein Sci 4 (8):1618–1632. doi:10.1002/pro.5560040820
10. Gromiha MM, Majumdar R, Ponnuswamy PK (1997) Identification of membrane spanning beta strands in bacterial porins. Protein Eng 10(5):497–500
11. Diederichs K, Freigang J, Umhau S et al (1998) Prediction by a neural network of outer membrane beta-strand protein topology. Protein Sci 7(11):2413–2420. doi:10.1002/pro.5560071119
12. Gromiha MM, Ahmad S, Suwa M (2004) Neural network-based prediction of transmembrane beta-strand segments in outer membrane proteins. J Comput Chem 25 (5):762–767. doi:10.1002/jcc.10386
13. Jacoboni I, Martelli PL, Fariselli P et al (2001) Prediction of the transmembrane regions of beta-barrel membrane proteins with a neural network-based predictor. Protein Sci 10 (4):779–787. doi:10.1110/ps.37201
14. Bagos PG, Liakopoulos TD, Spyropoulos IC et al (2004) A Hidden Markov Model method, capable of predicting and discriminating beta-barrel outer membrane proteins. BMC Bioinformatics 5:29. doi:10.1186/1471-2105-5-29
15. Bigelow HR, Petrey DS, Liu J et al (2004) Predicting transmembrane beta-barrels in proteomes. Nucleic Acids Res 32(8):2566–2577. doi:10.1093/nar/gkh580
16. Martelli PL, Fariselli P, Krogh A et al (2002) A sequence-profile-based HMM for predicting and discriminating beta barrel membrane proteins. Bioinformatics 18(Suppl 1):S46–S53
17. Park KJ, Gromiha MM, Horton P et al (2005) Discrimination of outer membrane proteins using support vector machines. Bioinformatics 21(23):4223–4229. doi:10.1093/bioinformatics/bti697
18. Garrow AG, Agnew A, Westhead DR (2005) TMB-Hunt: an amino acid composition based method to screen proteomes for beta-barrel transmembrane proteins. BMC Bioinformatics 6:56. doi:10.1186/1471-2105-6-56
19. Yan C, Hu J, Wang Y (2008) Discrimination of outer membrane proteins using a K-nearest neighbor method. Amino Acids 35(1):65–73. doi:10.1007/s00726-007-0628-7
20. Lin H (2008) The modified Mahalanobis Discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition. J Theor Biol 252(2):350–356. doi:10.1016/j.jtbi.2008.02.004
21. Ou YY, Gromiha MM, Chen SA et al (2008) TMBETADISC-RBF: discrimination of beta-barrel membrane proteins using RBF networks and PSSM profiles. Comput Biol Chem 32 (3):227–231. doi:10.1016/j.compbiolchem.2008.03.002
22. Fariselli P, Savojardo C, Martelli PL et al (2009) Grammatical-restrained hidden conditional random fields for bioinformatics applications. Algorithms Mol Biol 4:13. doi:10.1186/1748-7188-4-13
23. Hayat S, Elofsson A (2012) BOCTOPUS: improved topology prediction of transmembrane beta barrel proteins. Bioinformatics 28 (4):516–522. doi:10.1093/bioinformatics/btr710
24. Natt NK, Kaur H, Raghava GP (2004) Prediction of transmembrane regions of beta-barrel

proteins using ANN- and SVM-based methods. Proteins 56(1):11–18. doi:10.1002/prot.20092

25. Bagos PG, Liakopoulos TD, Hamodrakas SJ (2005) Evaluation of methods for predicting the topology of beta-barrel outer membrane proteins and a consensus prediction method. BMC Bioinformatics 6:7. doi:10.1186/1471-2105-6-7
26. von Heijne G (1992) Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. J Mol Biol 225 (2):487–494
27. Bannwarth M, Schulz GE (2003) The expression of outer membrane proteins for crystallization. Biochim Biophys Acta 1610(1):37–45, doi:S0005273602007113 [pii]
28. Pautsch A, Schulz GE (1998) Structure of the outer membrane protein A transmembrane domain. Nat Struct Biol 5(11):1013–1017. doi:10.1038/2983
29. Kozma D, Simon I, Tusnady GE (2013) PDBTM: Protein Data Bank of transmembrane proteins after 8 years. Nucleic Acids Res 41(Database issue):D524–D529. doi:10.1093/nar/gks1169
30. Delano WL (2002) The PyMOL molecular graphics system. http://www.pymol.org
31. Schulz GE (2002) The structure of bacterial outer membrane proteins. Biochim Biophys Acta 1565(2):308–317, doi: S0005273602005771 [pii]
32. Wimley WC (2002) Toward genomic identification of beta-barrel membrane proteins: composition and architecture of known structures. Protein Sci 11(2):301–312. doi:10.1110/ps.29402
33. Gromiha MM, Ponnuswamy PK (1993) Prediction of transmembrane beta-strands from hydrophobic characteristics of proteins. Int J Pept Protein Res 42(5):420–431
34. Zhai Y, Saier MH Jr (2002) The beta-barrel finder (BBF) program, allowing identification of outer membrane beta-barrel proteins encoded within prokaryotic genomes. Protein Sci 11(9):2196–2207. doi:10.1110/ps.0209002
35. Bishop CM, Walkenhorst WF, Wimley WC (2001) Folding of beta-sheets in membranes: specificity and promiscuity in peptide model systems. J Mol Biol 309(4):975–988. doi:10.1006/jmbi.2001.4715
36. Gnanasekaran TV, Peri S, Arockiasamy A et al (2000) Profiles from structure based sequence alignment of porins can identify beta stranded integral membrane proteins. Bioinformatics 16 (9):839–842
37. Freeman TC Jr, Wimley WC (2010) A highly accurate statistical approach for the prediction of transmembrane beta-barrels. Bioinformatics 26(16):1965–1974. doi:10.1093/bioinformatics/btq308
38. Liu Q, Zhu Y, Wang B et al (2003) Identification of beta-barrel membrane proteins based on amino acid composition properties and predicted secondary structure. Comput Biol Chem 27(3):355–361, doi: S1476927102000853 [pii]
39. Berven FS, Flikka K, Jensen HB et al (2004) BOMP: a program to predict integral beta-barrel outer membrane proteins encoded within genomes of Gram-negative bacteria. Nucleic Acids Res 32(Web Server issue): W394–W399. doi:10.1093/nar/gkh351
40. Altschul SF, Madden TL, Schaffer AA et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25(17):3389–3402, doi:gka562 [pii]
41. Bagos PG, Liakopoulos TD, Spyropoulos IC et al (2004) PRED-TMBB: a web server for predicting the topology of beta-barrel outer membrane proteins. Nucleic Acids Res 32 (Web Server issue):W400–W404. doi:10.1093/nar/gkh417
42. Randall A, Cheng J, Sweredoski M et al (2008) TMBpro: secondary structure, beta-contact and tertiary structure prediction of transmembrane beta-barrel proteins. Bioinformatics 24 (4):513–520. doi:10.1093/bioinformatics/btm548
43. Waldispuhl J, Berger B, Clote P et al (2006) transFold: a web server for predicting the structure and residue contacts of transmembrane beta-barrels. Nucleic Acids Res 34(Web Server issue):W189–193. doi:10.1093/nar/gkl205
44. Remmert M, Linke D, Lupas AN et al (2009) HHomp—prediction and classification of outer membrane proteins. Nucleic Acids Res 37(Web Server issue):W446–W451. doi:10.1093/nar/gkp325
45. Rabiner L (1989) A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE 77(2):257–286
46. Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14(9):755–763, doi: btb114 [pii]
47. Krogh A, Larsson B, von Heijne G et al (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305(3):567–580. doi:10.1006/jmbi.2000.4315

48. Nielsen H, Krogh A (1998) Prediction of signal peptides and signal anchors by a hidden Markov model. Proc Int Conf Intell Syst Mol Biol 6:122–130

49. Krogh A, Mian IS, Haussler D (1994) A hidden Markov model that finds genes in E. coli DNA. Nucleic Acids Res 22 (22):4768–4778

50. Krogh A (1994) Hidden Markov models for labelled sequences. In: Proceedings of the12th IAPR international conference on pattern recognition, pp 140–144

51. Chamberlain AK, Bowie JU (2004) Asymmetric amino acid compositions of transmembrane beta-strands. Protein Sci 13(8):2270–2274

52. Slusky JS, Dunbrack RL Jr (2013) Charge asymmetry in the proteins of the outer membrane. Bioinformatics 29(17):2122–2128. doi:10.1093/bioinformatics/btt355

53. Jackups R Jr, Liang J (2005) Interstrand pairing patterns in beta-barrel membrane proteins: the positive-outside rule, aromatic rescue, and strand registration prediction. J Mol Biol 354 (4):979–993. doi:10.1016/j.jmb.2005.09.094

54. Berman HM, Westbrook J, Feng Z et al (2000) The Protein Data Bank. Nucleic Acids Res 28 (1):235–242, doi:gkd090 [pii]

55. Andreeva A, Howorth D, Brenner SE et al (2004) SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res 32(Database issue): D226–D229. doi:10.1093/nar/gkh039

56. Lomize MA, Lomize AL, Pogozheva ID et al (2006) OPM: orientations of proteins in membranes database. Bioinformatics 22 (5):623–625. doi:10.1093/bioinformatics/btk023

57. Dobson L, Lango T, Remenyi I et al (2015) Expediting topology data gathering for the TOPDB database. Nucleic Acids Res 43(Database issue):D283–D289. doi:10.1093/nar/gku1119

58. Bagos PG, Tsaousis GN, Hamodrakas SJ (2009) How many 3D structures do we need to train a predictor? Genomics Proteomics Bioinformatics 7(3):128–137. doi:10.1016/S1672-0229(08)60041-8

59. Bagos PG, Hamodrakas SJ (2009) Bacterial beta-barrel outer membrane proteins: a common structural theme implicated in a wide variety of functional roles. In: Daskalaki A (ed) Handbook of research on systems biology applications in medicine, pp 182–207. doi:10.4018/978–1-60566-076-9.ch010

60. Punta M, Coggill PC, Eberhardt RY et al (2012) The Pfam protein families database. Nucleic Acids Res 40(Database issue): D290–D301. doi:10.1093/nar/gkr1065

61. Fariselli P, Finelli M, Marchignoli D et al (2003) MaxSubSeq: an algorithm for segment-length optimization. The case study of the transmembrane spanning segments. Bioinformatics 19(4):500–505

62. Zemla A, Venclovas C, Fidelis K et al (1999) A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. Proteins 34(2):220–223. doi:10.1002/(SICI)1097-0134(19990201)34:2<220::AID-PROT7>3.0.CO;2-K [pii]

63. Baldi P, Brunak S, Chauvin Y et al (2000) Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 16(5):412–424

64. Baum LE (1972) An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. Inequalities 3:1–8

65. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B Methodol 39(1):1–38. doi:10.2307/2984875

66. Krogh A (1997) Two methods for improving performance of an HMM and their application for gene finding. Proc Int Conf Intell Syst Mol Biol 5:179–186

67. Bagos P, Liakopoulos T, Hamodrakas S (2004) Faster gradient descent training of hidden Markov models, using individual learning rate adaptation. In: Paliouras G, Sakakibara Y (eds) Grammatical inference: algorithms and applications, vol 3264, Lecture notes in computer science. Springer, Berlin, Heidelberg, pp 40–52. doi:10.1007/978-3-540-30195-0_5

68. Krogh A, Riis SK (1999) Hidden neural networks. Neural Comput 11(2):541–563

69. Zou L, Wang Z, Wang Y et al (2010) Combined prediction of transmembrane topology and signal peptide of beta-barrel proteins: using a hidden Markov model and genetic algorithms. Comput Biol Med 40 (7):621–628. doi:10.1016/j.compbiomed.2010.04.006

70. Schwartz R, Chow YL (1990) The N-best algorithms: an efficient and exact procedure for finding the N most likely sentence hypotheses. In: 1990 international conference on acoustics, speech, and signal processing, 1990. ICASSP-90, 3–6 Apr 1990, vol 81, pp 81–84. doi:10.1109/icassp.1990.115542

71. Kall L, Krogh A, Sonnhammer EL (2005) An HMM posterior decoder for sequence feature prediction that includes homology

information. Bioinformatics 21(Suppl 1): i251–i257. doi:10.1093/bioinformatics/bti1014

72. Fariselli P, Martelli PL, Casadio R (2005) A new decoding algorithm for hidden Markov models improves the prediction of the topology of all-beta membrane proteins. BMC Bioinformatics 6(Suppl 4):S12

73. Won KJ, Hamelryck T, Prugel-Bennett A et al (2007) An evolutionary method for learning HMM structure: prediction of protein secondary structure. BMC Bioinformatics 8:357. doi:10.1186/1471-2105-8-357

74. Won KJ, Prugel-Bennett A, Krogh A (2004) Training HMM structure with genetic algorithm for biological sequence analysis. Bioinformatics 20(18):3613–3619. doi:10.1093/bioinformatics/bth454

75. Petersen TN, Brunak S, von Heijne G et al (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods 8(10):785–786. doi:10.1038/nmeth.1701

76. Lin K, Simossis VA, Taylor WR et al (2005) A simple and fast secondary structure prediction method using hidden neural networks. Bioinformatics 21(2):152–159. doi:10.1093/bioinformatics/bth487

77. Martelli PL, Fariselli P, Casadio R (2004) Prediction of disulfide-bonded cysteines in proteomes with a hidden neural network. Proteomics 4(6):1665–1671. doi:10.1002/pmic.200300745

78. Bagos PG, Liakopoulos TD, Hamodrakas SJ (2006) Algorithms for incorporating prior topological information in HMMs: application to transmembrane proteins. BMC Bioinformatics 7:189. doi:10.1186/1471-2105-7-189

79. Viklund H, Elofsson A (2004) Best alpha-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information. Protein Sci 13(7):1908–1917. doi:10.1110/ps.04625404