Bioinformatics Advance Access published November 15, 2007

Structural Bioinformatics

TMBpro: Secondary Structure, β-contact, and Tertiary Structure Prediction of Transmembrane β-Barrel Proteins

Arlo Randall[1,2], Jianlin Cheng[3], Michael Sweredoski[1,2], Pierre Baldi[1,2,*]

[1] School of Information and Computer Sciences, University of California, Irvine, CA, 92697.

[2] Institute for Genomics and Bioinformatics, University of California, Irvine, CA, 92697.

[3] Department of Computer Science, University of Missouri, Columbia, MO, 65203.

Associate Editor: Prof. Alfonso Valencia

ABSTRACT

Motivation: Transmembrane -barrel (TMB) proteins are embedded in the outer membranes of mitochondria, Gram-negative bacteria, and chloroplasts. These proteins perform critical functions, including active ion-transport and passive nutrient intake. Therefore there is a need for accurate prediction of secondary and tertiary structure of TMB proteins. Traditional homology modeling methods, however, fail on most TMB proteins since very few non-homologous TMB structures have been determined. Yet, because TMB structures conform to specific construction rules that restrict the conformational space drastically, it should be possible for methods that do not depend on target-template homology to be applied successfully.

Results: We develop a suite (TMBpro) of specialized predictors for predicting secondary structure (TMBpro-SS), β-contacts (TMBpro-CON), and tertiary structure (TMBpro-3D) of transmembrane -barrel proteins. We compare our results to the recent state-of-the-art predictors *transFold* and *PRED-TMBB* using their respective benchmark datasets, and leave-one-out-cross-validation. Using the *transFold* dataset TMBpro predicts secondary structure with per-residue accuracy (Q_2) of 77.8%, a correlation coefficient of .54, and TMBpro predicts β-contacts with precision of .65 and recall of .67. Using the *PRED-TMBB* dataset TMBpro predicts secondary structure with Q_2 of 88.3% and a correlation coefficient of .75. All of these performance results exceed previously published results by 4% or more. Working with the *PRED-TMBB* dataset, TMBpro predicts the tertiary structure of transmembrane segments with RMSD less than 6.0 Å for 9 of 14 proteins. For 6 of 14 predictions, the RMSD is less than 5.0 Å, with a GDT_TS score greater than 60.0.

Availability: http://www.igb.uci.edu/servers/psss.html

Contact: pfbaldi@ics.uci.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Transmembrane β-barrel (TMB) proteins are an important class of proteins embedded in the outer membrane of Gram-negative bacteria, mitochondria, and chloroplasts (Wallin and Heijne, 1998; Schulz, 2000; Tamm *et al.*, 2004). It is estimated that genomic databases currently contain thousands of TMB proteins (Wimley, 2002, 2003), and ongoing large-scale sequencing efforts promise to produce many more (Yooseph et al. 2007). These proteins carry out diverse biochemical functions including active ion transport, passive nutrient intake, and defense against attack proteins (Schulz, 2000; Koebnik *et al.*, 2000). Thus, elucidating the structure and function of TMB proteins has immediate medical relevance, as bacteria membrane proteins are potential targets of antimicrobial drugs and vaccines (Jackups and Liang, 2005). Crystallizing transmembrane (TM) proteins is especially challenging; thus, predicting the structure of TMB proteins from sequence is an interesting and important task (Casadio *et al.*, 2003; Oberai *et al.*, 2006).

Currently, several methods try to discriminate TMB proteins from globular and TM β-helical proteins, or to predict their 1-dimensional (1D) secondary structure features (i.e., the positions of TM β-strands and the types of loops) (Paul and Rosenbusch, 1985; Welte *et al.*, 1991; Gromiha *et al.*, 1997; Diederichs *et al.*, 1998; Jacoboni *et al.*, 2001; Martelli *et al.*, 2002; Zhai and Saier, 2002; Liu *et al.*, 2003; Bagos *et al.*, 2004*a,* 2004*b*; Bigelow *et al.*, 2004; Gromiha *et al.*, 2004; Natt *et al.*, 2004; Bagos *et al.*, 2005; Fariselli *et al.*, 2005; Gromiha and Suwa, 2005; Gromiha *et al.*, 2005; Garrow *et al.*, 2005; Park *et al.*, 2005; Bigelow and Rost, 2006; Waldispühl *et al.*, 2006*b*).

The 1D structure predictions are very useful for constructing a coarse topology of TMB structure (Tamm *et al.*, 2001). However, they do not provide enough information to construct a low-resolution tertiary structure for a TMB protein (Jackups and Liang, 2005). In addition, traditional homology modeling of TMB proteins is hindered by the lack of sequence similarity between the small number of TMB proteins with known structures and the thousands of TMB proteins without known structures (Schulz, 2000; Jacoboni *et al.*, 2001).

TMB proteins adopt a common β-barrel fold and obey specific construction rules, as outlined in (Schulz, 2000). For instance, known TMB proteins consist of an even number of membrane spanning β-strands with an anti-parallel β-meander topology. Two recently published methods take advantage of these construction rules to predict the inter-strand β-residue pairings of TMB proteins (Jackups and Liang, 2005; Waldispühl *et al.*, 2006*a*). These β-contact predictions provide strong constraints for building tertiary structure models of TMB proteins as in the reconstruction of globular protein structures using contact constraints (Skolnick *et al.*, 1997).

Since there are fewer than 20 non-redundant (Waldispühl *et al.*, 2006*a*) TMB proteins with known structures in the Protein Data Bank (PDB) (Berman *et al.*, 2000) and membrane protein databases (Ikeda *et al.*, 2003; Lomize *et al.*, 2006), it is challenging to develop robust knowledge-based methods to predict inter-strand pairings in TMB proteins. To overcome the small dataset problem the method *transFOLD* (Waldispühl *et al.*, 2006*a*) uses pair-wise inter-strand residue statistical potentials derived from globular proteins to predict the inter-strand residue pairings of TMB proteins with moderate accuracy.

In this paper, we present a three-stage pipeline to predict the tertiary structure of TMB proteins. First, we predict the two-class secondary structure with TMBpro-SS. Second, we predict β-residue contacts using TMBpro-CON (Pollastri and Baldi, 2002; Baldi and Pollastri, 2003; Cheng *et al.*, 2006*a*; Cheng and Baldi, 2005). Finally, we use these feature predictions, TMB templates, and construction rules to predict tertiary structure TMBpro-3D.

2 DATA

2.1 Benchmark sets

In this work we use two sets of TMB proteins described in the literature. The first is the dataset described in (Waldispühl *et al.*, 2006*a*), which consists of 14 redundancy-reduced TMB proteins. The authors divide this set into two main subsets: non-water-filled (NWF) and water-filled (WF). NWF consists of (PDB code) 1QJP, 1QJ8, 1THQ, 1P4T, 1I78, 1K24, 1QD6. WF consists of 1A0S, 1AF6, 1PRN, 2OMF, 1E54, 1TLY, and 2POR. In our work, we treat all 14 proteins as a single set. The secondary structure assignments used for this set come from the DSSP program (Kabsch and Sander, 1983), which we condense to two classes: strand (β) and non-strand (-). These single character designations are used throughout this work when dealing with two-class representation. Following the work described in (Waldispühl *et al.*, 2006*a*), the group published a web-server for predicting features of TMB proteins called *transFold* (Waldispühl *et al.*, 2006*b*). Throughout this work, we refer to this set as *SetTransfold*. We compare our secondary structure and β-contact prediction results to *transFold* using this set.

The second set is described in (Bagos *et al.*, 2004*a*) and also contains 14 redundancy-reduced TMBs. Nine of them overlap with *SetTransfold*: 1QJP, 1QJ8, 1I78, 1K24, 1A0S, 1PRN, 2OMF, 1E54, and 2POR. The five proteins that differ are: 1QD5, 2MPR, 1FEP, 2KMO, and 2FCP. Rather than using the DSSP assignments, the authors manually designated TM (β) and non-TM (-) segments for each protein in this set. This approach was motivated by the observation that many of the β-strands in TMB proteins extend significantly beyond the membrane, and the authors sought to focus on the TM regions. The authors have made their method available as the web server *PRED-TMBB* (Bagos *et al.*, 2004*b*). For the remainder of this work we refer to this set as *SetPRED-TMBB*. We compare our results for secondary structure and topology prediction to *PRED-TMBB* using this set. We also use this set to evaluate our tertiary structure predictions.

The two datasets are created and treated independently in this work in order to make fair comparisons to previous work. For all of the proteins in *SetTransFold* the secondary structure annotation comes from DSSP. For all of the proteins in *SetPRED-TMBB* the secondary structure, the annotation comes from manual designation. For the nine proteins common to both datasets we keep both types of secondary structure annotation. For example, protein 1QJ8 is present in each dataset, but with different secondary structure annotation (DSSP in *SetTransFold* and manual designation in *SetPRED-TMBB*). Results comparing our work to *transFold* are based only on *SetTransFold* annotations, and results comparing our work to *PRED-TMBB* are based solely on *SetPRED-TMBB* annotations.

We compare our results using sets *SetTransFold* and *SetPRED-TMBB* to the published results of the respective methods. To compare our β-contact predictions to those of *transFold* using the same predicted secondary structure, we submitted the proteins of *SetTransFold* to the *transFold* server. The *transFold* server predicts the secondary structure into four classes: membrane facing strand residues (M), channel facing strand residues (C), loops inside the periplasm (i), and extra-cellular loops (o). *TransFold* also predicts β-residue contacts. These single character designations are used throughout this work and in the output of our server. The *PRED-TMBB* server predicts secondary structure into three classes: TM, periplasmic, and extra-cellular. For both datasets we expanded the two-class representation to three-class by designating 'β' residues as either 'M' or 'C' based on visual inspection of the structures. These representations (M, C, -) were used to train a three-class predictor.

2.2 Cross-validation

Our predictors are trained and tested using leave-one-out cross-validation (LOOCV) on *SetTransFold* and *SetPRED-TMBB* independently. A single protein is held out of the set, a model is built using the other thirteen, and a prediction is made on the held out protein. This process is repeated for each protein in the set to obtain the evaluation statistics in the results section. LOOCV provides the best estimate of the generalization accuracy of a predictor; however, with larger datasets LOOCV is not practical because of the training time involved in building a model for each member of the dataset. The same LOOCV procedure is applied to template usage in the tertiary structure prediction evaluation. The procedure is also commonly referred to as 'Jackknife'.

2.3 Template construction

Our tertiary prediction evaluation is performed using *SetPRED-TMBB*. We created template files by extracting the backbone (N, Cα, C) coordinates from the monomeric PDB files. The curated (β, -) designations are used to label each residue position in the template. The set contains 2 proteins with 8 strands, 2 with 10 strands, 1 with 12 strands (1QD5), 4 with 16 strands, 2 with 18 strands, and 3 with 22 strands. The strand count of the predicted secondary structure is used to select templates for modeling. If the strand count of 1QD5 is correctly predicted, no templates would be available for modeling because of the LOOCV procedure. To account for this, we built a template from one additional 12 stranded protein: 1TLY. Also, if a 14 stranded protein is predicted, no templates would be available; therefore, we built templates from two 14 stranded TMBs: 1T16 and 2F1C. The manually curated designations were not available for these three proteins, so we used the TM segment ranges published in the Orientation of Proteins in Membranes (OPM) database (Lomize *et al.*, 2006). The template set contains no 20 stranded proteins because none are present in the PDB.

3 METHODS

3.1 Secondary structure prediction

3.1.1 Neural-network implementation The TMB secondary structure predictor uses specialized neural network architecture called a 1-Dimensional Recursive Neural Network (1D-RNN). This network architecture has been used for prediction of secondary structure, SSpro (Pollastri *et al.*, 2002), domain boundaries, DOMpro (Cheng *et al.*, 2006*b*), and disordered regions, DISpro (Cheng *et al.* 2005). As in the prior applications, the input at each position to the neural network is the profile of the sequences in the NR database aligned to the target sequence using PSI-BLAST (Altschul *et al.*, 1997). It has been the experience of the authors that there is little chance of over-fitting the models because of the weight sharing involved in the 1D-RNN architecture. This feature of the architecture makes it appropriate for the small datasets used in this work.

3.1.2 Two-class prediction (β, -) For two-class prediction the 1D-RNN is trained on the two-class 1D representation: (β) and (-). When making a prediction, the output from the model is the predicted probability of class membership to each class. The initial predicted secondary structure, $S_{initial}$, consists of the class with higher predicted probability at each position. The first row in Figure 1 contains an example of $S_{initial}$ for the TMB protein 1P4T. Since the secondary structure of TMB proteins adhere to consistent construction rules, we perform post-processing on the predicted probabilities to revise the secondary structure prediction. The lengths of β-segments and the different types of loop segments are constrained by minimum and maximum values; however, the length of N and C-terminal (-) segments are left unconstrained. Table S1 in the Supplementary Materials contains a summary of the specific values used for the different segment types for each dataset. In the example in Figure 1, the initial secondary structure

SS Source	Predicted or Assigned Secondary Structure of Protein 1P4T
Initial Pred 2($S_{initial}$)	-----EEEEEEEEEEE-E--------EEEEEEEEEEEEEEEEEEEEEEEE-E-----EEEEEEEEEEEEEEE-------E--EEEEEEEEEEEE-----EEEEEEEE-EEEEEEEEEE-E-EEEEEEEEEEE------EEEEEEEEEEEEEEEE-
Pred 2 (S_{max})	-----EEEEEEEEEEEEE--------EEEEEEEEEEE--EEEEEEEEEEE-------EEEEEEEEEEEEEEE-------EEEEEEEEEEEEEEE-----EEEEEEEEEEEEEEEEEEE---EEEEEEEEEEE------EEEEEEEEEEEEEEEE-
Pred 4	.....MCMCMCMCMCMCMooooooooMCMCMCMCMCMiiMCMCMCMCMCMoooooooMCMCMCMCMCMCMCMiiiiiiiMCMCMCMCMCMCMCMoooooMCMCMCMCMCMCMCMCMCMiiiMCMCMCMCMCMooooooMCMCMCMCMCMCMCMC.
Annotation	...MCMCMCMCMCMCMCMooooMCMCMCMCMCMCMCMiiMCMCMCMCMCMCooooooooMCMCMCMCMCMCMiiiiiMCMCMCMCMCMCMCMCMooooCMCMCMCMCMCMCMCMCMCMMiiMCMCMCMCMCMCMMCooCMMCMCMCMCMCMCMC.

Fig. 1. Predicted secondary structure for protein 1P4T. LOOCV prediction made using *SetTransfold*. *Initial Pred 2 ($S_{initial}$)* is the initial two-class prediction by the neural network. *Pred 2 (S_{max})* is the two-class prediction after post-processing. *Pred 4* is the four class prediction with loop types inferred from *Pred 2 (S_{max})* and membrane/channel pattern predicted by the three-class predictor. *Annotation* is the 1D sequence according to the DSSP designations for strand boundaries and our assignment of 'M', 'C', 'i', 'o', and '.' based on visual inspection.

prediction $S_{initial}$ for protein 1P4T violates multiple constraints. To describe the post-processing strategy formally we use the additional notations: N is the number of residues in a sequence, S is any two-class secondary structure that does not violate any of the model constraints, S_i is the secondary structure at position i, O is the matrix of predicted probabilities output from the 1D-RNN, $O_{i,\beta}$ and $O_{i,non\text{-}\beta}$ are the predicted probabilities that S_i is 'β' or '-', respectively. The post-processing objective function is the sum of predicted probabilities for each position of S as defined in Equation 1.

$$sum(S) = \sum_{i=1}^{N} O_{i,S_i} \qquad (1)$$

Given *sum(S)* as the objective function, we need to find an S which maximizes *sum(S)*, which we denote S_{max}. If $S_{initial}$ does not violate any of the constraints, then no search is necessary as $sum(S_{max}) \leq sum(S_{initial})$. To find an S_{max} we developed a dynamic-programming (DP) solution that incorporates the parameters of the TMB construction rules. The search guarantees to find an S_{max}, but the solution may not be unique. Since, we have no objective way to discriminate between two equal scoring predictions this issue is ignored, and the single optimal path returned from the DP search is used as the final S_{max}.

We use the number of β-strands in S_{max} as the prediction of strand count. During the search for S_{max} the DP method saves the value of *sum(S)* for each value of potential strand count. If the number of strands 'θ' is provided as an additional constraint, the notation $S_{max,\theta}$ indicates an optimal S with θ strands. This information can be useful for assessing the confidence in the predicted secondary structure and corresponding strand count. Table S2 in the Supplementary Materials contains a summary of the $S_{max,\theta}$ results for the proteins in *SetPRED-TMBB*. For 1QJ8 the gap between $S_{max,8}$ (130.4) and the next highest sum $S_{max,10}$ (115.2) is 11.7%, whereas for 1A0S the gap between $S_{max,16}$ (340.9) and the next highest sum $S_{max,18}$ (340.1) is only 0.2%. The larger the gap, the more confident the predictor is in its strand count. For assessing our system, this information is not useful, as the predictor will use the single best S_{max}; however, this information could be valuable to a user who may decide to build tertiary models from multiple strand counts.

3.1.3 Three-class prediction (M, C, -) To predict the membrane/channel pattern within the β segments we trained a separate neural network to predict three classes: M, C, and other (-). The architecture for the three-class predictor is the same 1D-RNN architecture used for the two-class predictor. The output of the network is the probability of class membership in each of the three classes. For each β segment predicted in the final two-class prediction S_{max}, the membrane-channel (M/C) pattern is predicted by choosing the pattern with the higher predicted probability sum. For the example protein, 1P4T, in Figure 1, the first β segment is predicted to be from position 6 to 18. Equation 2 shows the calculation for the sum of predicted probabilities for each pattern.

$$\begin{aligned} sum\_MC &= O_{6,M} + O_{7,C} + \ldots + O_{17,C} + O_{18,M} \\ sum\_CM &= O_{6,C} + O_{7,M} + \ldots + O_{17,M} + O_{18,C} \end{aligned} \qquad (2)$$

In this case *sum_MC* > *sum_CM* so the pattern beginning with 'M' is forced over the β segment. From the three-class prediction, the (-) segments are assigned as periplasmic (i) or extra-cellular (o) according to the pattern observed in all TMB proteins. See Figure 1 for the final four-class prediction of the example protein 1P4T in comparison to the annotations.

3.2 β-contact prediction

Between two paired β-strands, only every other pair of aligned residues is hydrogen bonded. Residue pairs that are aligned, but not hydrogen bonded to one another, are still considered β-contacts. The DSSP program is used to automatically identify β-contacts in known protein structures. DSSP classifies β-contacts based on inter-residue atomic distances and angles. TMBpro-CON is trained on true β-contacts using a 2 Dimensional Recursive Neural Network (2D-RNN) (Cheng and Baldi, 2005). TMBpro-CON predicts β-contacts in TMB proteins by first predicting the probability of pairing between all pairs of predicted β-strand residues. For each pair of strands the pseudo-energy (i.e. the sum of the individual predicted pairing probabilities) of all possible strand-strand alignments is calculated. Then TMBpro-CON utilizes the following rules to restrict the search for acceptable pairings: consecutive strands must pair in anti-parallel fashion; the terminal strands must pair in anti-parallel fashion; the shear number must be between 0 and +4 with respect to the strand count; membrane facing residues must pair with other membrane facing residues; and core facing residues must pair with other core facing residues. A dynamic programming method is used to find a set of contact predictions that maximizes the global pseudo-energy while conforming to the construction rules.

3.3 Tertiary structure prediction

TMBpro-3D combines *de novo* and template based methods to predict tertiary structure, using a search energy composed of predicted structural feature, physical interaction, and statistical terms. The conformational search is performed using simulated annealing with a move set that utilizes whole protein templates and fragment assembly.

3.3.1 Search energy The search energy used in the conformational search is a linear combination of the following terms:

- E_{beta_pairs} - favors hydrogen bonding between predicted β-contacts.

- $E_{mc\_pattern}$ - favors predicted M/C pattern using template residue membrane-channel values.
- $E_{globular\_pairwise}$ - rewards favorable side-chain interactions between predicted non-β positions (Zhang *et al.*, 2003).
- E_{chain_break} - favors close termini proximity at artificial chain break sites.
- $E_{centroid\_repulsion}$ - penalizes clashes between side-chain centers of mass.
- $E_{vdw\_repulsion}$ - penalizes steric clashes between all explicitly modeled atoms using Van der Waals radii.

The details of each individual and the corresponding weights are provided in the Supplementary Materials.

3.3.2 Template usage The strand count (θ) of the predicted secondary structure is used to screen for potential templates. Each template with a strand count matching θ is used to generate an ensemble of models. All models are then ranking according to their energy, and the model with the best search energy is the final tertiary prediction. To allow flexible alignment of each predicted β-segment to its corresponding template segment, TMBpro creates artificial chain breaks at the center of each non-β region, dividing the model into θ loosely coupled sub-models. The sub-models are allowed to move independently, but their interactions are captured through the global energy function.

Four arrays of variables ($\mathcal{M}$, $\mathcal{T}$, $\mathcal{U}$, $\mathcal{H}$) are used to manage template utilization during the conformational search (see Figures 2 and S2). The model $\mathcal{M}$ is an array containing the xyz coordinates of the backbone atoms (N, C_α, C), indexed by the residue number i. The template $\mathcal{T}$ is a similar array built for the template protein. The template usage $\mathcal{U}$ is an array of binary variables indicating whether or not $\mathcal{T}$ is used to model $\mathcal{M}$ at each residue position. $\mathcal{U}_i$=1 indicates that $\mathcal{T}$ is used to model $\mathcal{M}_i$, while $\mathcal{U}_i$=0 means $\mathcal{M}_i$ is modeled by fragment replacement using the fragment library (Simons *et al.*, 1997). The alignment shifts $\mathcal{H}$ is an array of length θ, where each position is the integer shift between model and template segment relative to center-center alignment. Initially the centers of all model and template segments are aligned, corresponding to $\mathcal{H}_i = 0$ for i=1,…,θ. From these center-center alignments, $\mathcal{U}$ is set to 1 at each predicted β position that aligns to a β-residue in the template, and the rest of $\mathcal{U}$ is set to 0 (Figures 2 and S2). During the search phase the values of $\mathcal{H}$ and $\mathcal{U}$ are modified to explore the use of $\mathcal{T}$.

3.3.3 Move types The following move types are used in the simulated annealing protocol to search the conformational space:

- *Shift Single Segment by k*: $\mathcal{H}_i = \mathcal{H}_i + k$
  i = segment index; $k \in \mathbf{Z}$ and $-max \le k \le max$;
  max = (length of segment i)/2;
- *Shift m Consecutive Segments by k*: $\mathcal{H}_j = \mathcal{H}_j + k$, for $j=i,\ldots,i+m-1$
  i = starting segment index; $m \in \mathbf{Z}$ and $2 \le m \le \theta$;
  $k \in \mathbf{Z}$ and $-max \le k \le max$;
  max = (length of shortest among m segments)/2;
- *Adjust Single Segment Template Usage by k*: $\mathcal{U}_l = \delta$ for $l=b,\ldots,b+k-1$
  b = index of boundary residue ($\mathcal{U}_b \ne \mathcal{U}_{b+1}$);
  $\delta =$ 0 (contraction) or $\delta =$ 1 (extension);
  $k \in \mathbf{Z}$ and $-max \le k \le max$;
  max = number of residues to next boundary;
- *Replace with Fragment*: use fragment to model $\mathcal{M}_i,\ldots\mathcal{M}_{i+k}$
  i = index of first residue to replace;
  $k \in \mathbf{Z}$ and $1 \le k \le 9$;
  This move is applied only to regions where the template is not used ($\mathcal{U}_i,\ldots,\mathcal{U}_{i+k}$=0).

3.3.4 Conformational search The space of possible conformations is searched using simulated annealing with a linear cooling schedule and the move-set described above. The search is performed in two distinct phases. Phase 1 focuses on modeling the TM-segments, while phase 2 focuses on modeling the loops. In phase 1 all move types are used and the weights for $E_{globular\_pairwise}$, E_{chain_break}, $E_{centroid\_repulsion}$, and $E_{vdw\_repulsion}$ are set to 0 to allow the search to quickly find a conformation that satisfies the predicted strand constraints (low E_{beta_pairs} and $E_{mc\_pattern}$). At the end of phase 1 the values of $\mathcal{H}$ are locked, so that the model-template alignments are no longer allowed to change. This reduces the move set in phase 2 to only *Adjusting Single Segment Template Usage* and *Replace with Fragment*. In addition, all energy terms are used in phase 2. The search is run with different random seeds to generate an ensemble of predicted models, equally utilizing the available templates. The model with the lowest final search energy is returned as the tertiary structure prediction.

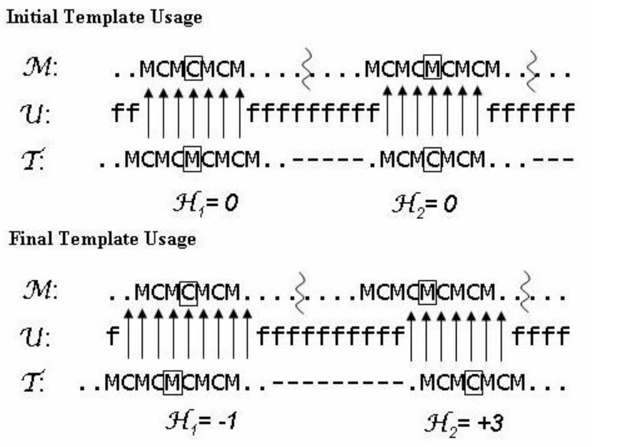
Initial Template Usage
M: ..MCMCMCM....⌇...MCMCMCMCM..⌇..
U: ff↑↑↑↑↑↑↑fffffffff↑↑↑↑↑↑↑ffffff
T: ..MCMCMCMCM..-----.MCMCMCM...---
H1= 0
H2= 0
Final Template Usage
M: ..MCMCMCM....⌇...MCMCMCMCM..⌇..
U: f↑↑↑↑↑↑↑↑↑fffffffff↑↑↑↑↑↑↑ffff
T: ..MCMCMCMCM..---------.MCMCMCM...
H1= -1
H2= +3

Fig. 2. Hypothetical template usage example for the first two TM segments. $\mathcal{M}$ and $\mathcal{T}$ represent the model and template respectively. $\mathcal{U}$ controls where the template is used: '↑' indicates the position is modeled from $\mathcal{T}$ ($\mathcal{U}_i$ = 1), whereas '**f**' means the position is modeled by fragment modeling ($\mathcal{U}_i$ = 0). The wavy vertical lines mark the chain breaks. The center residue of each segment is boxed to help illustrate the shifts. Initially the centers of segments are aligned (all $\mathcal{H}_i$ = 0). In the final model, the 1st segment is shifted 1 position to the left ($\mathcal{H}_1$ = -1) and the 2nd segment is shifted 3 positions to the right ($\mathcal{H}_2$ = 3).

4 RESULTS

To assess our secondary structure prediction we compare it to the published results of *transFold* (Waldispühl *et al.*, 2006*a*) and *PRED-TMBB* (Bagos *et al.,* 2004*a*) methods. To assess our β-contact prediction we compare it to the published results of *transFold*, and to the server output in order to make a comparison using the same predicted secondary structure as input. To the best of our knowledge, TMBpro-3D is the first publicly available method to predict the structure of TMB proteins without relying on sequence-sequence, sequence-profile, or profile-profile alignments for template usage; thus, we do not compare out tertiary prediction results to previous work.

4.1 Secondary structure prediction results

As described previously, we developed a two-class (β,-) secondary structure predictor specialized for TMB proteins. Using the two-class predictions we predict the three-class (M, C, -) and infer four-class predictions (M, C, i, o). We developed two separate secondary structure predictors using the non-redundant datasets

Table 1. TMBpro-SS compared to *transFold*

Method	Q_2	MCC	SOV	Q_3	$Q_\beta^{\%\,obs}$	$Q_\beta^{\%\,pred}$
transFold	69.9	.380	--	58.5	94.9	85.2
TMBpro-SS	77.8	.538	.800	71.5	97.2	88.2

For this comparison TMBpro-SS is evaluated using LOOCV on the *SetTransfold* dataset. Comparison metrics are: Q_2: two-class per-residue accuracy, MCC: Mathews correlation coefficient, SOV: segment overlap measure, Q_3: three-class per-residue accuracy $Q_\beta^{\%obs}$: per-segment recall (sensitivity), $Q_\beta^{\%pred}$: per-segment precision.

SetTransfold and *SetPRED-TMBB* to make comparisons with the related methods.

4.1.1 Secondary structure evaluation metrics To assess secondary structure prediction performance we use the following per-residue metrics: the two-class accuracy (Q_2), three-class (M, C, -) accuracy (Q_3), Mathews Correlation Coefficient (MCC) (Baldi *et al*., 2000), and Segment Overlap Measure (SOV) (Zemla *et al.,* 1999).We include the SOV measure for completeness, but no SOV results were provided in the studies we compare to. In addition to these common measures, we use additional measures from previous work for the sake of comparison. For comparison to *transFold* we also include the per-segment recall (sensitivity) $Q_\beta^{\%obs}$ and precision $Q_\beta^{\%pred}$, with correct prediction defined as an observed β-strand intersecting exactly one predicted β-strand, and vice versa (Waldispuhl *et al.,* 2006*a*). The per-segment measures for comparison to *PRED-TMBB* include the number of true positives (TP), the number of false negatives (FN), and the number of false positives (FP). In addition, we include the number of correctly predicted topologies (TOP), that is when all strands and loops have been predicted correctly according to (Bagos *et al.,* 2004*a*).

4.1.2 Results using SetTransfold Table 1 contains a summary of TMBpro-SS secondary structure prediction results compared to *transFold*. We use LOOCV on *SetTransfold* to assess our method and compare it to *transFold*. TMBpro-SS outperforms *transFold* significantly using the Q_2 (77.84% to 69.91%) and MCC (.538 to .380) measures. TMBpro-SS performs slightly better than *transFold*, according to the per-segment measures $Q_\beta^{\%obs}$ and $Q_\beta^{\%pred}$.

4.1.3 Results using SetPRED-TMBB Table 2 contains a summary of TMBpro-SS secondary structure prediction results compared to *PRED-TMBB*. We use the same LOOCV (Jackknife) procedure as the authors of the *PRED-TMBB* method, on the same set of proteins, to make the comparison as objective as possible. Of the 214 annotated β-strands PRED-TMBB correctly predicts 203, while TMBpro-SS correctly predicts 204. PRED-TMBB makes 13 false positive predictions (FP), while TMBpro-SS only makes 6. Using the TOP measure of correct topology prediction PRED-TMBB correctly predicts 8 topologies, while TMBpro-SS succeeds on 11. TMBpro-SS also outperforms PRED-TMBB according to the Q_2 (88.3% to 84.2%) and MCC (.751 to .720) measures. When comparing TMBpro-SS to itself between datasets it has significantly higher Q_2, Q_3, MCC and SOV when using *SetPRED-TMBB* (see Tables 1 and 2). It is unclear how much of this difference is due to the five proteins that differ between the sets, and how much is due to the different types of annotation of the training data. The Q_2, Q_3, MCC and SOV results for individual proteins are displayed with the detailed tertiary prediction results in Table 4.

Table 2. Secondary structure prediction compared to *PRED-TMBB*

Method	TP	FP	FN	TOP	Q_2	Q_3	MCC	SOV
PRED-TMBB	203	13	11	8	84.2	--	.720	--
TMBpro-SS	204	6	10	11	88.3	88.0	.751	91.3

TMBpro-SS is evaluated using LOOCV on the *SetPRED-TMBB* dataset and compared to *PRED-TMBB*. Per-segment measures are; TP: true positives, FP: false positives, FN: false negatives. Topology measure; TOP: correct topology. Per-residue measures; Q_2: two-class accuracy, Q_3: three-class accuracy, MCC: Mathews correlation coefficient and SOV: segment overlap measure.

Table 3. β-contact prediction results.

Dataset / Method	Precision δ=0	Recall δ=0	Precision δ=2	Recall δ=2
SetTransfold				
transFold – published	--	--	.350	.450
transFold – server results	.084	.105	.434	.512
TMBpro-CON (*transFold*)	.110	.128	.445	.532
TMBpro-CON (TMBpro-SS)	.206	.215	.648	.671
TMBpro-CON (DSSP)	.478	.520	.960	.960
SetPRED-TMBB				
TMBpro-CON (TMBpro-SS)	.414	.407	.851	.819
TMBpro-CON (annotation)	.484	.529	.967	.996

Summary of β-contact prediction results. The secondary structure method used by TMBpro-CON is in parentheses. For δ=0 only exact pairs are counted, for δ=2 pairings within ± 2 are counted as correct. True β-contacts are determined by the DSSP program.

4.2 β-contact prediction results

The input to TMBpro-CON is the amino acid sequence and a two-class secondary structure. Using *SetTransfold* we performed β-contact prediction with three different sets of two-class secondary structure: (1) predicted by *transFold* server, (2) predicted by TMBpro-SS and (3) DSSP designations. We compare our results using (1) to the β-contacts predicted by the *transFold* server. We compare our results using (2) to the *transFold* published results. Using *SetPRED-TMBB* we performed β-contact prediction with two sets of two-class secondary structure: predicted by TMBpro-SS and hand curated annotations from (Bagos *et al*., 2004*a*). No comparison to other work is made using *SetPRED-TMBB* since *PRED-TMBB* does not predict β-contacts.

4.2.1 β-contact evaluation metrics For evaluation of β-contact prediction the authors of *transFold* introduce the concept of a *compatible pair* of residues to allow contact predictions that are nearly correct to be counted. Consider a pair (*i,j*) to be a true β-

Table 4. TMBpro-3D tertiary structure prediction results.

secondary structure prediction results							tertiary prediction results (predicted SS and β-contacts)					self-consistency results (curated SS and β-contacts)				
PDB ID	**Length**	**TM count**	**MCC**	**SOV**	Q_2	Q_3	**β-recall δ=0**	**GDT_TS**	**RMSD**	**GDT_TS $_{TM}$**	**RMSD $_{TM}$**	**GDT_TS**	**RMSD**	**GDT_TS $_{TM}$**	**RMSD $_{TM}$**	**RMSD $_{TM\ SelfTemplate}$**
1QJ8	148	8	.86	97.3	93.2	93.2	.64	52.0	5.5	69.9	3.6	58.8	5.2	76.5	2.3	0.0
1QJP	171	8	.78	95.0	89.5	89.5	.65	57.3	4.3	68.3	3.0	54.2	4.9	63.7	2.8	0.0
1K24	253	10	.59	86.7	80.2	80.2	.18	32.3	12.9	50.0	10.2	55.7	4.8	76.9	1.9	0.0
1QD5	269	12	.68	95.2	84.4	79.6	.37	25.5	11.7	37.1	8.3	41.1	8.6	62.3	4.5	0.0
1PRN	289	16	.81	91.9	90.3	89.6	.46	50.0	7.1	68.0	5.7	55.6	5.4	76.5	2.0	0.0
1I78	297	10	.89	99.7	95.3	95.3	.66	35.9	14.8	66.1	4.0	41.2	14.5	79.3	1.7	0.0
2POR	301	16	.57	70.6	78.7	77.1	.28	29.7	13.4	43.8	11.4	58.7	5.4	81.9	1.5	0.0
1E54	332	16	.73	95.5	86.7	85.8	.48	49.3	7.7	70.9	4.4	55.5	7.1	79.5	2.7	0.0
2OMF	340	16	.84	97.7	92.4	90.6	.31	41.8	8.6	66.3	4.9	54.5	5.9	81.8	1.8	0.0
1A0S	413	16(18)	.5	65.2	75.3	74.3	.17	21.9	16.8	33.5	14.1	66.9	5.1	89.8	1.7	0.0
2MPR	427	16(18)	.53	76.7	77.5	76.6	.41	29.4	13.7	40.2	12.3	67.9	7.7	92.3	1.8	0.0
2FCP	723	22	.85	98.7	93.9	93.9	.39	25.9	15.5	48.8	6.0	41.5	13.9	74.9	3.5	0.0
1FEP	724	22	.81	97.4	91.7	91.7	.54	38.7	11.0	60.2	4.4	48.0	10.0	78.7	2.7	0.0
1KMO	741	22	.88	99.2	95.1	95.1	.34	31.1	9.1	53.3	5.3	54.6	8.7	78.9	2.1	0.0

TM count: number of transmembrane segments in secondary structure predicted by TMBpro-SS. If the prediction does not match the true number of segments, the true number is shown in parentheses. MCC: Mathews correlation coefficient. SOV: segment overlap measure (Zemla *et al.,* 1999). Q_2: two-class accuracy. Q_3: three-class accuracy. β-recall δ=0: recall of exact hydrogen bonded β-residue pairs. GDT_TS: global distance test total score. RMSD: root-mean-squared deviation. The TM notation indicates the assessment was only performed on the portions of the protein annotated as transmembrane. The results in section 'tertiary prediction results' are generated using predicted secondary structure and β-contacts. The results in section 'self-consistency results' are generated using the manually curated secondary structure of *SetPRED-TMBB* and true β-contacts as determined by DSSP. The final column in the self-consistency section, RMSD_{TM}SelfTemplate, shows results when TMBpro is allowed to use all available templates (including the self template), all other results are generated using LOOCV template selection.

residue pairing. The contact pairs (*i,j*) and (*m,n*) are considered to be compatible if, for a given integer δ, (*i,j*) = (*m* ± δ, *n* ± δ). In their work they use a value of δ=2 for evaluation. For our assessment we use δ=2 and δ=0, where only exact pairing predictions are counted. The measures we use for assessment are precision and recall. The precision is calculated by (number of correct β-contact predictions / total number of β-contact predictions) and recall by (number of correct β-contact predictions / total number of true β-contacts).

4.2.2 Results using SetTransfold A summary of β-contact prediction results for both protein sets and all secondary structure sets is available in Table 3. Using the same secondary structure as input (the predicted secondary structure from the *transFold* server) TMBpro-CON performs slightly better than the *transFold* server by all measures. Using the predicted secondary structure from TMBpro-SS as input, TMBpro-CON performs significantly better than *transFold* server results and published results according to all measures. Using the DSSP assigned secondary structure as input TMBpro-CON predicts exact β-contacts with precision .478 and recall .520. These results demonstrate the upper bound in β-contact prediction accuracy of TMBpro-CON given improvements in secondary structure prediction only.

4.2.3 Results using SetPRED-TMBB Taking the predicted secondary structure from TMBpro-SS trained on *SetPRED-TMBB* as input, TMBpro-CON predicts exact β-contacts with precision .414 and recall .407. These values are significantly higher than the corresponding prediction using *SetTransfold* (see Table 3). This difference can be accounted for by the more accurate secondary structure predictions for *SetPRED-TMBB*. The β-contact recall results for the individual proteins are shown in the tertiary results Table 4.

4.3 Tertiary structure prediction results

Here we evaluate the tertiary structure predictions of TMBpro-3D for *SetPRED-TMBB* using secondary structure and β-contacts predicted by TMBpro. We chose *SetPRED-TMBB* rather than *SetTransfold* for tertiary prediction experiments because of the stronger secondary structure and β-contact prediction results. Only the model with the lowest search energy is evaluated.

4.3.1 Tertiary structure evaluation metrics The two measures we use to evaluation tertiary predictions are root-mean-square deviation (RMSD) and global distance test total score (GDT_TS). The latter has been used as the primary numeric measure in recent critical assessment of methods of protein structure prediction (CASP) experiments (Zemla *et al*., 2001). The TM notation is used as a subscript to indicate that the measure is calculated on only the TM segments of the true structure compared to the model.

4.3.2 Prediction results The tertiary structure prediction results for each protein in *SetPRED-TMBB* are displayed in Table 4. The best prediction, in terms of the GDT_TS and RMSD on the whole structure is made on the protein with the second highest β-contact recall: 1QJP. The β-contact recall is .65, the GDT_TS is 57.3 and RMSD is 4.3 Å. The $\text{GDT\_TS}_{\text{TM}}$ is 68.3 and RMSD_{TM} is 3.0 Å. The next best whole structure predictions are for proteins 1QJ8 (52.0, 5.5 Å), 1PRN (50.0, 7.1 Å), and 1E54 (49.3, 7.7 Å). The Supplementary Materials contains a superposition file (1QJ8_pred.pdb) and an image (Figure S1) showing the predicted structure for 1QJ8 aligned to the PDB structure. For several proteins the $\text{GDT\_TS}_{\text{TM}}$ results are strong. For proteins 1QJ8, 1QJP, 1PRN, 1I78, 1E54, 2OMF and 1FEP the $\text{GDT\_TS}_{\text{TM}}$ is greater than 60.0. These predictions correspond to correct topology predictions and high β-contact recall when compared to the other predictions. The significantly lower GDT_TS and higher RMSD scores on the whole structures reflect the difficulty of modeling long loop regions and core domains folded inside the larger proteins.

The worst whole structure and TM segment predictions are made on proteins 1A0S and 2MPR, both of which have true strand counts of 18, but are modeled using 16-stranded templates because of incorrect secondary structure topology predictions. Additionally, the locations of multiple strands in the 2POR prediction are incorrect resulting in an incorrect topology according to the TOP measure. The worst whole structure and TM segment prediction for a protein with correct topology prediction was made on the 10-stranded protein 1K24. The topology is correct using the TOP measure; however, the locations of the sixth and seventh strands are off by seven residues. Using a slightly stricter standard for topology assessment, this prediction would be considered an incorrect topology. From these results it is clear that the correct topology is necessary to build a reasonable tertiary model.

4.3.3 Self-consistency results To evaluate the self-consistency of TMBpro we provided the curated secondary structure and true β-contacts as input to the program. The performance was assessed both allowing and disallowing the inclusion of the native template among the available templates, and the results are displayed in the rightmost section of Table 4. When the native template is included, TMBpro always recovers the true structure (see the last column in Table 4). When the native template is not included, the RMSD_{TM} results range from 1.5 Å to 4.5 Å. For 12 of 14 predictions, the RMSD_{TM} is less than 2.8 Å. The only two exceptions are proteins 2FCP, with an RMSD_{TM} of 3.5 Å, and 1QD5, with an RMSD_{TM} of 4.5 Å. At 723 residues 2FCP is one of the longest proteins in the set, so a slightly higher error is not surprising. 1QD5 is only 269 residues, but contains an irregular bulge in the first strand that is not present in its only available template (1TLY).

5 CONCLUSION

TMB proteins have clear biological and medical relevance. Due to their importance and the difficulty of experimentally determining their structures, accurate tertiary structure prediction of TMB proteins is an important task for the protein structure prediction community. Traditional homology modeling methods will perform well if the target protein is similar enough to a solved protein to create a quality alignment; however, for the vast majority of putative TMB proteins traditional homology modeling will fail. The construction rules TMB proteins follow provide a greatly reduced search space compared to the globular protein structure prediction problem. In this work we demonstrated a methodology for predicting secondary structure, β-contacts, and tertiary structure of TMB proteins. The tertiary structure predictor does not rely on sequence similarity between target and template. The performance of TMBpro compares favorably to other publicly available predictors. The TMBpro server, trained on all 14 proteins in *SetPRED-TMBB*, is publicly available at: http://www.igb.uci.edu/servers/psss.html.

ACKNOWLEDGEMENTS

Work supported by NIH grant LM-07443-01, NSF grants EIA-0321390 and IIS-0513376, and a Microsoft Faculty Research Award to PB, and a UCF faculty start-up grant to JC.

REFERENCES

Altschul,S., Madden,T., Schaffer,A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389-3402.

Bagos,P., Liakopoulos,T. and Hamodrakas,S. (2005) Evaluation of methods for predicting the topology of beta-barrel outer membrane proteins and a consensus prediction method. *BMC Bioinformatics,* **6**, 7.

Bagos,P., Liakopoulos,T., Spyropoulos,I. and Hamodrakas,S. (2004*a*) A hidden markov model method, capable of predicting and discriminating beta-barrel outer membrane proteins. *BMC Bioinformatics,* **5**, 29.

Bagos,P., Liakopoulos,T., Spyropoulos,I. and Hamodrakas,S. (2004*b*) PRED-TMBB: a web server for predicting the topology of beta-barrel outer membrane proteins. *Nucleic Acids Res.,* **32**, W400–W404.

Baldi,P., Brunak,S., Chauvin,Y., Andersen,C. and Nielsen,H. (2000) Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, **16**, 412-424.

Baldi,P. and Pollastri,G. (2003) The principled design of large-scale recursive neural-network architectures-DAG-RNNs and the protein structure prediction problem. *Journal of Machine Learning Research,* **4**, 575–602.

Berman,H., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T., Weissig,H., Shindyalov,I. and Bourne,P. (2000) The Protein Data Bank. *Nucleic Acids Res.,* **28**, 235–242.

Bigelow,H., Petrey,D., Liu,J., Przybylski,D. and Rost,B. (2004) Predicting transmembrane beta-barrels in proteomes. *Nucleic Acids Res.,* **32**, 2566–2577.

Bigelow,H. and Rost,B. (2006) PROFtmb: a web server for predicting bacterial transmembrane beta barrel proteins. *Nucleic Acids Res.,* **34**, W186–W188.

Casadio,R., Fariselli,P. and Martelli,P. (2003) In silico prediction of the structure of membrane proteins: is it feasible? *Brief. Bioinformatics,* **4**, 341–348.

Cheng,J. and Baldi,P. (2005) Three-stage prediction of protein beta-sheets by neural networks, alignments, and graph algorithms. *Bioinformatics,* **21** (Suppl. 1), i75–i84.

Cheng,J., Saigo,H. and Baldi,P. (2006*a*) Large-scale prediction of disulphide bridges using kernel methods, two-dimensional recursive neural networks, and weighted graph matching. *Proteins,* **62** (3), 617–629.

Cheng,J., Sweredoski,M. and Baldi,P. (2005) Accurate Prediction of Protein Disordered Regions by Mining Protein Structure Data. *Data Mining and Knowledge Discovery*, **11** (3), 213-222.

Cheng,J., Sweredoski,M. and Baldi,P. (2006*b*) DOMpro: protein domain prediction using profiles, secondary structure, relative solvent accessibility, and recursive neural networks. *Data Mining and Knowledge Discovery*, **13** (1), 1-10.

Diederichs,K., Freigang,J., Umhau,S., Zeth,K. and Breed,J. (1998) Prediction by a neural network of outer membrane beta-strand protein topology. *Protein Sci.,* **7** (11), 2413–2420.

Fariselli,P., Martelli,P. and 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**, S12.

Garrow,A., Agnew,A. and Westhead,D. (2005) TMB-Hunt: an amino acid composition based method to screen proteomes for beta-barrel transmembrane proteins. *BMC Bioinformatics,* **6**, 56.

Gromiha,M., Ahmad,S. and Suwa,M. (2005) TMBETA-NET: discrimination and prediction of membrane spanning beta-strands in outer membrane proteins. *Nucleic Acids Res.,* **33**, W164–W167.

Gromiha,M., Ahmad,S. and Suwa,M. (2004) Neural network-based prediction of transmembrane beta-strand segments in outer membrane proteins. *J. Comput. Chem.,* **25**, 762–767.

Gromiha,M., Majumdar,R. and Ponnuswamy,P. (1997) Identification of membrane spanning beta strands in bacterial porins. *Protein Engineering,* **10**, 497–500.

Gromiha,M. and Suwa,M. (2005) A simple statistical method for discriminating outer membrane proteins with better accuracy. *Bioinformatics,* **21**, 961–968.

Ikeda,M., Arai,M., Okuno,T. and Shimizu,T. (2003) TMPDB: a database of experimentally-characterized transmembrane topologies. *Nucleic Acids Res.,* **31**, 406–409.

Jackups,R. and 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**, 979–993.

Jacoboni,I., Martelli,P., Fariselli,P., Pinto,V.D. and Casadio,R. (2001) Prediction of the transmembrane regions of beta-barrel membrane proteins with a neural network based predictor. *Protein Sci.,* **10** (4), 779–787.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22** (12), 2577-2637.

Koebnik,R., Locher,K. and Gelder,P.V. (2000) Structure and function of bacterial outer membrane proteins: barrels in a nutshell. *Mol. Microbiol.,* **37**, 239–253.

Liu,Q., Zhu,Y., Wang,B. and Li,Y. (2003) A HMM-based method to predict the transmembrane regions of beta-barrel membrane proteins. *Comput. Biol. Chem.,* **27**, 69–76.

Lomize,M., Lomize,A., Pogozheva,I. and Mosberg,H. (2006) OPM: orientations of proteins in membrane database. *Bioinformatics,* **22**, 623–625.

Martelli,P., Fariselli,P., Krogh,A. and Casadio,R. (2002) A sequence-profile-based hmm for predicting and discriminating beta barrel membrane proteins. *Bioinformatics,* **18**, S46–S53.

Moult,J., Krzysztof,F., Rost,B., Hubbard,T. and Tramontano,A. (2005) Critical assessment of methods of protein structure prediction (CASP) - Round 6. *Proteins*, **61** (Suppl. 7), 3-7.

Natt,N., Kaur,H. and Raghava, G. (2004) Prediction of transmembrane regions of beta-barrel proteins using ANN- and SVM-based methods. *Proteins*, **56**, 11-18.

Oberai,A., Ihm,Y., Kim,S. and Bowie,J. (2006) A limited universe of membrane protein families and folds. *Protein Sci.,* **15**, 1723–1734.

Park,K., Gromiha,M., Horton,P. and Suwa,M. (2005) Discrimination of outer membrane proteins using support vector machines. *Bioinformatics,* **21**, 4223–4229.

Paul,C. and Rosenbusch,J. (1985) Folding patterns of porin and bacteriorhodopsin. *EMBO J.,* **4**, 1593–1597.

Pollastri,G. and Baldi,P. (2002) Prediction of contact maps by GIOHMMs and recurrent neural networks using lateral propagation from all four cardinal corners. *Bioinformatics,* **18**, S62–S70.

Pollastri,G., Przybylski,D., Rost,B., and Baldi,P. (2002) Improving the Prediction of Protein Secondary Structure in Three and Eight Classes Using Recurrent Neural Networks and Profiles. *Proteins*, **47**, 228-235.

Schulz,G. (2000) Beta-barrel membrane proteins. *Curr. Opin. Struct. Biol.,* **10**, 443–447.

Simons,K.T., Kooperberg,C., Huang,E. and Baker,D. (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.*, **268**, 209-225.

Skolnick,J., Kolinski,A. and Ortiz,A. (1997) Monsster: a method for folding globular proteins with a small number of distance restraints. *J. Mol. Biol.,* **265**, 217–241.

Tamm,L., Arora,A. and Kleinschmidt,J. (2001) Structure and assembly of beta-barrel membrane proteins. *J. Biol. Chem.,* **276**, 32399–32402.

Tamm,L., Hong,H. and Liang,B. (2004) Folding and assembly of beta barrel membrane proteins. *Biochim. Biophys. Acta,* **1666**, 250–263.

Waldispühl,J., Berger,B., Clote,P. and Steyaert,J. (2006*a*) Predicting transmembrane beta-barrels and interstrand residue interactions from sequence. *Proteins,* **65** (1), 61–74.

Waldispühl,J., Berger,B., Clote,P. and Steyaert,J. (2006*b*) transFold: a web server for predicting the structure and residue contacts of transmembrane beta-barrels. *Nucleic Acids Res.,* **34**, W189–W193.

Wallin,E. and von Heijne,G. (1998) Genome wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci.,* **7** (4), 1029–1038.

Welte,W., Weiss,M., Nestel,U., Weckesser,J., Schiltz,E. and Schulz,G. (1991) Prediction of the general structure of OmpF and PhoE from the sequence and structure of porin from Rhodobacter capsulatus. Orientation of porin in the membrane. *Biochim. Biophys. Acta,* **1080** (3), 271–274.

Wimley,W. (2002) Toward genomic identification of beta-barrel membrane proteins: composition and architecture of known structures. *Protein Sci.,* **11**, 301–312.

Wimley,W. (2003) The versatile beta-barrel membrane protein. *Curr. Opin. Struct. Biol.,* **13**, 404–411.

Yooseph,S., Sutton,G., Rusch,D.B., Halpern,A.L., Williamson,S.J., Remington,K., Eisen,J.A., Heidelberg,K.B., Manning,G., Li,W., Jaroszewski,L., Cieplak,P., Miller,C.S., Li,H., Mashiyama,S.T., Joachimiak,M.P., van Belle,C., Chandonia,J.M., Soergel,D.A., Zhai,Y., Natarajan,K., Lee,S., Raphael,B.J., Bafna,V., Friedman,R., Brenner,S.E., Godzik,A., Eisenberg,D., Dixon,J.E., Taylor,S.S., Strausberg,R.L., Frazier,M. and Venter,J.C. (2007) The Sorcerer II Global Ocean Sampling expedition:cexpanding the universe of protein families. *PLoS Biol*, 5 (3), e16.

Zhai,Y. and Saier,M. (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.

Zhang,T., Kolinski,A. and Skolnick,J. (2003) TOUCHSTONE:II a new approach to ab initio protein structure prediction. *Biophys. J.,* **85**, 1145-1164.

Zemla,A., Venclovas,C., Fidelis,K. and Rost,B. (1999) A modified definition of sov, a segment-based measure for protein secondary structure prediction assessment. *Proteins*, **34**, 220-223.

Zemla,A., Venclovas,C., Moult,J. and Fidelis,K. (2001) Processing and evaluation of predictions in CASP4. *Proteins*, **45** (Suppl. 5), 13-21.