BIOINFORMATICS

Vol. 28 ISMB 2012, pages i90–i96
doi:10.1093/bioinformatics/bts233

Ranking models of transmembrane β-barrel proteins using Z-coordinate predictions

Sikander Hayat and Arne Elofsson*

Center for Biomembrane Research, Department of Biochemistry and Biophysics,
Stockholm Bioinformatics Center, Science for Life Laboratory, Swedish E-science Research Center, Stockholm University, SE-10691 Stockholm, Sweden

ABSTRACT

Motivation: Transmembrane β-barrels exist in the outer membrane of gram-negative bacteria as well as in chloroplast and mitochondria. They are often involved in transport processes and are promising antimicrobial drug targets. Structures of only a few β-barrel protein families are known. Therefore, a method that could automatically generate such models would be valuable. The symmetrical arrangement of the barrels suggests that an approach based on idealized geometries may be successful.

Results: Here, we present tobmodel; a method for generating 3D models of β-barrel transmembrane proteins. First, alternative topologies are obtained from the BOCTOPUS topology predictor. Thereafter, several 3D models are constructed by using different angles of the β-sheets. Finally, the best model is selected based on agreement with a novel predictor, ZPRED3, which predicts the distance from the center of the membrane for each residue, i.e. the Z-coordinate. The Z-coordinate prediction has an average error of 1.61 Å. Tobmodel predicts the correct topology for 75% of the proteins in the dataset which is a slight improvement over BOCTOPUS alone. More importantly, however, tobmodel provides a $C\alpha$ template with an average RMSD of 7.24 Å from the native structure.

Availability: Tobmodel is freely available as a web server at: http://tobmodel.cbr.su.se/. The datasets used for training and evaluations are also available from this site.

Contact: arne@bioinfo.se

1 INTRODUCTION

There are two classes of integral transmembrane membrane proteins, α-helical proteins and β-barrels. α-helical transmembrane proteins constitute 20–30% of a typical genome. The transmembrane β-barrels, which are the focus of this article, are less abundant and only found in the outer membrane of gram-negative bacteria, chloroplast and mitochondria. Although less abundant, transmembrane β-barrels are known to play a crucial role in the transport over the membrane and, additionally, in pore formation. They are also candidate targets for the development of antimicrobial drugs (Galdiero *et al.*, 2007; Koebnik *et al.*, 2000; Pajón *et al.*, 2006; Schulz, 2002). The experimental determination of the structure of transmembrane proteins is fraught with difficulties and, thus, computational methods are essential for identification and structural prediction. In particular, such predictions can be used for further experimental investigations and might aid in elucidating the function of these proteins.

Given the symmetrical structure of β-barrels it is possible to generate models directly from a theoretical description of a barrel (Chou *et al.*, 1990; Murzin *et al.*, 1994a,b). Obviously, these structural models might provide insights into the interactions between residues. Further, a 3D model can be used to design site-directed mutagenesis experiments and can be used in the modeling of large membrane protein complexes, such as the TOM-complex of the mitochondrial outer membrane (Becker *et al.*, 2012).

Many methods have been developed for the prediction of β-barrel transmembrane protein topology (Bagos *et al.*, 2004, 2005; Bigelow and Rost, 2006; Freeman and Wimley, 2010; Gromiha *et al.*, 2004, 2005; Martelli *et al.*, 2002; Mirus and Schleiff, 2005; Remmert *et al.*, 2009; Singh *et al.*, 2011; Wimley, 2002; Yan *et al.*, 2011). In addition, methods like partiFold (Waldispühl *et al.*, 2008) and TMBpro (Randall *et al.*, 2008) predict the inter β-strand residue contacts. However, to our knowledge only two methods predict a full 3D model, TMBpro (Randall *et al.*, 2008) and 3D-SPoT (Naveed *et al.*, 2012). In fact, the latter requires the protein topology as input, so it is only TMBpro that provides the full sequence to structure prediction.

TMBpro is based on secondary structure predictions and prediction of residue contacts in potential transmembrane β-barrels. TMBpro uses templates derived from known transmembrane β-barrels and aligns the predicted secondary structure to one of the 18 predefined templates based on the number of predicted β-strands. However, as discussed by Naveed *et al.*, template-based methods cannot be applied to novel folds, such as transmembrane β-barrels with an odd number of β-strands (Naveed *et al.*, 2012). 3D-SPoT is based on optimizing hydrogen bonds and side chain interactions between adjacent β-strands. Three different types of bonds are taken into account; strong hydrogen bonds, weak hydrogen bonds and side chain interactions. During optimization, adjacent β-strands are shifted up and down to obtain the lowest energy arrangement (Naveed *et al.*, 2012). The final model is then built using an intertwined coil geometric model, where each β-strand is represented by a coil wrapped around a hypothetical cylinder, and each coil is separately modeled. Finally, main chain atoms are added to the $C\alpha$ trace using an algorithm developed by Gront *et al.* (2007).

Here, we present tobmodel—a computational method for modeling of transmembrane β-barrels. The workflow of tobmodel is outlined in Figure 1. As shown, first BOCTOPUS is used to generate a set of potential topologies for a given protein sequence (Hayat and Elofsson, 2012). The second step is to generate a set of alternative 3D models using β-barrels with idealized geometries for each topology. Finally, to rank the models, we developed a novel

*To whom correspondence should be addressed.

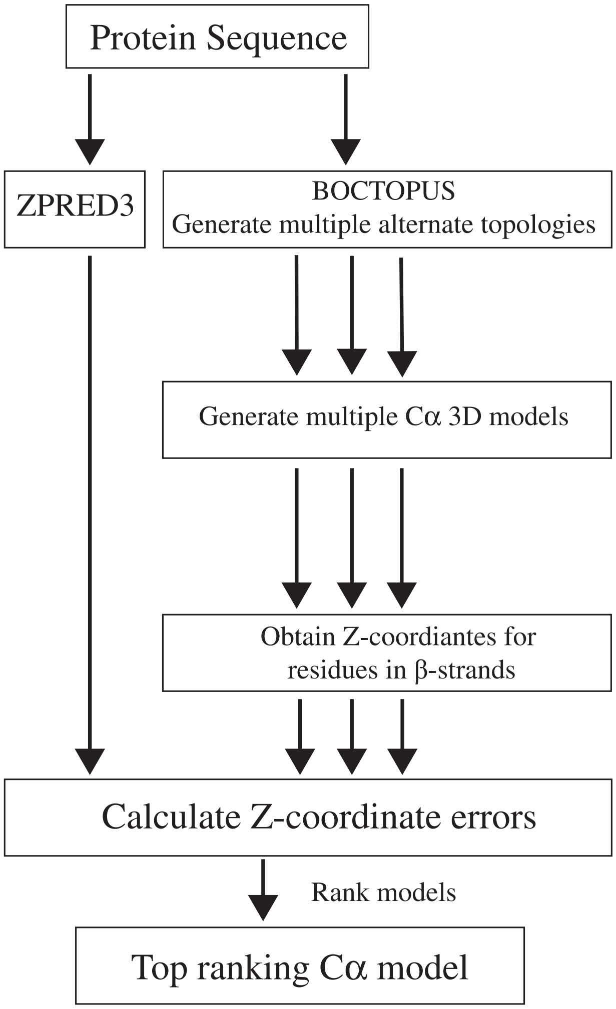

Fig. 1. Tobmodel pipeline. First, BOCTOPUS (Hayat and Elofsson, 2012) is used to generate topologies. Thereafter, 3D $C\alpha$ models are generated using the predicted topologies and different shear numbers. Finally, ZPRED3 is used to predict the Z-coordinate of each residue, then the models are ranked based on the lowest Z-coordinate error

'Z-coordinate' predictor, ZPRED3. The ranking is determined by the agreement between, on the one hand, the Z-coordinate predictions from ZPRED3 and, on the other, the Z-coordinates from the model. Further, tobmodel provides better topologies than TMBpro, based on a 10-fold cross-validation test. The quality of the final model generated by tobmodel is comparable to the final TMBpro model, while 3D-SPoT seems to generate better models.

2 METHODS

2.1 Training dataset

The dataset contains 36 transmembrane β-barrel proteins obtained from OPM (Lomize *et al.*, 2006) with ≤30% sequence identity to each other. All training was based on a strict 10-fold cross-validation. Furthermore, to avoid homology bias all proteins belonging to the same OPM family were put together in the same cross-validation. For comparison purposes all proteins in the datasets used in TMBpro and 3D-SPoT were also downloaded from OPM. All datasets are available on the tobmodel web server.

2.2 Topology prediction using BOCTOPUS

BOCTOPUS is a two-stage topology prediction method for transmembrane β-barrels recently developed by Hayat *et al.* (Hayat and Elofsson, 2012). The predictor consists of two stages that take the local and global preferences of each residue into account. In the first stage, three separate support vector machines are used to predict the local preference of each residue to be either in the membrane (M), the inner-loop (i) or the outer-loop (o). Here, position specific scoring matrixes (PSSMs) are employed as the input to the support vector machines. Thereafter, the probability of a residue to be in each of these regions is used to create the profile. The second stage consists of a Hidden Markov Model (HMM), representing transmembrane β-barrels. The profile generated from the first stage is used as the input to the second stage. The HMM contains states for short inner loops, up and down strands and long outer loops. The start and end of the protein sequence are normally fixed to be at the periplasmic side, i.e. at the inner loop, and therefore only even number of β-strands are produced. This is in agreement with the known structural characteristics of bacterial outer membrane β-barrels. However, for β-barrels in mitochondria and chloroplasts an odd number of strands can also exist.

Emission and transition probabilities are not trained by the BOCTOPUS HMM. Instead all emission parameters are set to 1 and so are most transition parameters. However, a few transition parameters had to be optimized. As a result, during the training and 10-fold cross-validation of BOCTOPUS, a range of parameters were found to provide similar accuracy in terms of correctly predicted topology. Therefore, BOCTOPUS might produce a large number of alternative topologies for some proteins while for other proteins, where the predictions are more reliable, all these parameters produce the same topology. Here, we use this feature to generate a number of alternative topologies for a given protein sequence.

2.3 ZPRED3

ZPRED and ZPRED2 predict the Z-coordinate, i.e. the distance from the membrane center to a residue, for α-helical membrane proteins (Granseth *et al.*, 2006; Papaloukas *et al.*, 2008) using artificial neural networks. Here, we extend the ZPRED approach to also predict the Z-coordinate of β-barrel proteins, and replaced the artificial neural networks with support vector machines.

As when developing ZPRED, the target function was set to be the absolute value of Z-coordinate given an upper and a lower limit. For β-barrel proteins we found suitable upper and lower limits to be 15 and 3 Å. Several different input features were tested during the development of ZPRED3, (Table 1). These include the following; a PSSM, 20-bit sparse encoding of amino acids (AASparse), 3-bit encoded predicted topology where each bit represents inner loop, membrane and outer loop regions for the predicted topology obtained from BOCTOPUS (Topology), probabilities obtained from the SVM stage of BOCTOPUS (SVMDATA) and combinations of these features. For PSSM generation we used PSI-BLAST (Altschul *et al.*, 1997) with default parameters and three iterations of searching against the non-redundant nr-database, obtained from the NCBI website in July 2010. The log-odds value in the PSSM was transformed into a PSSM-profile by dividing all numbers by 10, such that they lie between ±1.0, as in BOCTOPUS (Hayat and Elofsson, 2012).

For the training dataset, structures aligned to the membrane normal were obtained from the OPM database (Lomize *et al.*, 2006). The task of training an SVM is then to predict the distance from the center of the membrane. Similar to the development of BOCTOPUS, the dataset was divided into 10 sets, such that proteins belonging to the same super family were in the same set. During training, nine sets were used to test the performance on the 10th set. Different SVMs, as implemented in the libsvm interface in the R e1071 package, were trained to predict the Z-coordinate. We tested radial basis and linear kernels, and different windows sizes in the range of 1–21. For each set of input variables, the optimal window size was determined based on the highest correlation coefficient and percentage of residues predicted to be within 2 Å of their observed Z-coordinate.

Table 1. Z-coordinate prediction comparison

Inputs	ws1	ws2	CC	Accuracy	Z-error	Q2 (%)
PSSM	7	–	0.72	60%	2.18	63
PSSM+Topology	1	19	0.78	74%	1.61	71
AASparse	5	–	0.45	44%	3.23	37
AASparse+PSSM	1	3	0.57	51%	2.75	49
SVMDATA	21	–	0.74	72%	1.78	68
Topology	21	–	0.75	73%	1.71	69
Topology+AASparse	17	1	0.74	72%	1.75	68

Development of ZPRED3 and selection of input variables. Here, ws1 and ws2 refer to the window size of the input feature window. CC is the correlation coefficient between the observed and the predicted Z-coordinate. Accuracy refers to the percentage of residues found to be within 2 Å of their observed Z-coordinate. Z-error refers to the average Z-coordinate error. Q2 is the number of residues correctly predicted to be within the membrane region. For all sets of inputs, window sizes and kernels were optimized. Only the best performing set of parameters for each set of input variables is shown here.

The Z-coordinate prediction accuracy of ZPRED3 was finally evaluated based on (i) correlation coefficient, (ii) number of residues correctly predicted to be within 2 Å of their observed distance from the membrane center, (iii) average error per protein and (iv) correctly identified membrane residues (Q2). Here, Q2 is defined as a two-state accuracy measure that accounts for the number of residues that are predicted to be within the membrane/non-membrane region. The best results were obtained using PSSM and a window size of 1, and the 3-bit encoded predicted topology obtained from BOCTOPUS with a window size of 19 (Table 1).

2.4 Modeling of TM regions of β-barrels based on predicted topology and theoretical principles

Here, the 3D modeling of residues predicted to be in transmembrane β-strands is based on the geometric optimization method for ideal barrels as previously described (Chou *et al.*, 1990; Murzin *et al.*, 1994a,b). In short, a single-chain transmembrane β-barrel forms a closed cylindrical barrel. An ideal β-barrel can be defined using five parameters, (i) the number of strands (N), (ii) the shear number (S), (iii) the barrel radius (R), (iv) the strand tilt (T) and (v) the twist and coil of the β-strand. In all known transmembrane β-barrels, the β-strands are right-twisted, and in order to satisfy hydrogen bonding, each β-strand is right-tilted with respect to the membrane normal. The shear number S, is a measure of the staggering of the β-strands and can be described as follows.

Choose a residue r_i on any β-strand from the β-barrel. Starting from this residue r_i, if we follow the residues on adjacent strands that form a hydrogen bond with our chosen residue in a counter-clockwise direction, then we will not arrive at the initially chosen residue r_i but at residue r_i+S, where S is the shear number. A positive and negative value of S signifies a right and left-tilted barrel, respectively.

Transmembrane β-barrels consist of a right-handed barrel that can be approximated by a circle of radius R, than can be theoretically estimated using Equation 1 (Murzin *et al.*, 1994a,b).

$$R=\frac{[(Sa)^2+(Nb)^2]^{1/2}}{2N\pi\sin(\pi/n)} \tag{1}$$

Where a is the $C\alpha-C\alpha$ distance between two residues along a strand (typically 3.3 Å), b is the interstrand distance, 4.4Å (Murzin *et al.*, 1994b).

The tilt T is defined as the slope of the strands to the axis of the barrel. Based on the number of strands N and the shear number S, the tilt T of the transmembrane β-strands can be calculated using Equation (2).

$$\tan(T)=\frac{Sa}{Nb} \tag{2}$$

Where a and b are defined as in Equation (1).

We calculate a range of values for R and T based on the number of predicted strands, N, found in a topology and shear numbers in the range of 6–24. The parameters describing the twist and the coil of the β-strand are not optimized in the current study but is an area of future development. However, this should only slightly affect the vertical distance of β-strand residues from the membrane center, i.e. the effect of this on the selection process by ZPRED3 should only be marginal.

The β-strand can then be transformed from the Cartesian coordinate system to a β-strand coordinate system defined by three axis (f, g, h) as described in Chou *et al.* (1984). This transformation is performed to make the translation and rotation of β-sheets relative to each other more convenient. Briefly, the Cartesian coordinate system has the origin at (x = y = z = 0) and the three axes are perpendicular to each other. In a β-strand coordinate system, one β-strand is chosen as the reference and the h axis is defined as the axis of this β-strand. The origin of the (f, g, h) coordinate system is chosen as the midpoint of the reference β-strand. The other β-strands are then transformed from the Cartesian coordinate system to the β-strand coordinate system. Transformation of the coordinate system from (x, y, z) to (f, g, h) is given by Equation (3):

$$\begin{bmatrix} f \\ g \\ h \end{bmatrix}=l_s\begin{bmatrix} x-x_{\dot{s}} \\ y-y_{\dot{s}} \\ z-z_{\dot{s}} \end{bmatrix} \tag{3}$$

Where $x_{\dot{s}}$, $y_{\dot{s}}$ and $z_{\dot{s}}$ are the coordinates of residues in the reference strand and l_S is defined as

$$l_S=\begin{bmatrix} (e_f)_x & (e_f)_y & (e_f)_z \\ (e_g)_x & (e_g)_y & (e_g)_z \\ (e_h)_x & (e_h)_y & (e_h)_z \end{bmatrix} \tag{4}$$

Where, $(e_f)_x$ denotes the projection of unit vector (e_f) on the x axis, etc.

Here, e_h defines the direction of the h axis in a Cartesian coordinate system (x, y, z), and is given by the equation:

$$e_h=le_x+me_y+ne_x \tag{5}$$

Where e_x, e_y and e_z are unit vectors pointing along the coordinate axes x, y, and z, respectively. l, m, n are the direction cosines of the reference strand in a general (x, y, z) coordinate system.

The direction of the f axis of the β-strand coordinate system is defined by the unit vector e_f, given by:

$$e_f=\frac{e_h\times\overrightarrow{C_i^\alpha C_j^\alpha}}{\left|e_h\times\overrightarrow{C_i^\alpha C_j^\alpha}\right|} \tag{6}$$

Where $\overrightarrow{C_i^\alpha C_j^\alpha}$ is a vector pointing from the i-th to the j-th $C\alpha$ atom in a β-strand in the x, y, z coordinate system. Here, i and j can be any pair of residues, preferably in the middle of the β-strand.

Finally, the direction of the g axis is defined by the unit vector e_g, and is given by

$$e_g=e_h\times e_f \tag{7}$$

The reverse transformation from (f, g, h) to (x, y, z) coordinate system is given by Equation (8):

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix}={l_S}^{-1}\begin{bmatrix} f \\ g \\ h \end{bmatrix}+\begin{bmatrix} x_{\dot{s}} \\ y_{\dot{s}} \\ z_{\dot{s}} \end{bmatrix} \tag{8}$$

2.5 Ranking of $C\alpha$ models

Two Z-coordinates were obtained for each residue in each model β-barrel, one predicted Z-coordinate was provided by the ZPRED3 prediction and one Z-coordinate was calculated from the model. The $C\alpha$ models were then ranked according to the average difference between these two Z-coordinates.

Table 2. Method comparison

Methods	Correct topology	Average RMSD	Average TM_Score
TMBpro[a]	19 (36)	8.79	0.56
tobmodel[a]	27 (36)	7.24	0.43
BOCTOPUS[a]	25.4±2.0 (36)	NA[b]	NA[b]
3D-SPoT[c]	NA (23)[d]	4.10[d]	–[d]
tobmodel[c]	18 (23)	7.06	0.43
tobmodel[c,e]	23 (23)	5.86	0.48

Comparison of models generated by different prediction methods. TMBpro employs predefined templates extracted from proteins in their dataset.

[a] tobmodel data.

[b] BOCTOPUS does not generate 3D models.

[c] 3D-SPoT dataset.

[d] 3D-SPoT does not test alternative topologies and the results are taken as reported previously (Naveed *et al.*, 2012).

[e] For comparison, we have here used the correct topologies into the tobmodel pipeline.

3 RESULTS AND DISCUSSION

Here, we present a pipeline to model transmembrane β-barrel proteins. In short, we use BOCTOPUS (Hayat and Elofsson, 2012) to predict a large set of alternative topologies for a protein. Subsequently, for each topology, we generate a large number of possible models using theoretical descriptions of β-barrel proteins. Finally, we predict the 'Z-coordinate' for each residue using a novel predictor, ZPRED3, and use the agreement between these predictions and the models for ranking.

3.1 Topology predictions

Given our dataset of 36 proteins, BOCTOPUS consistently predicts the incorrect topology of four proteins (3PRN_C, 1E54_E, 2VQI_A and 3CSL_A), while for 16 proteins, the correct topologies are consistently predicted (Hayat and Elofsson, 2012). However, BOCTOPUS produces a mixture of correct and incorrect topologies of 20 proteins. Among the 20 proteins with mixed topologies, tobmodel identifies the correct topology for 11, while BOCTOPUS on average predicts the correct topologies for 9.4.

In summary, the topology prediction accuracy of BOCTOPUS alone is 25.4 ± 2.0, while tobmodel predicts 27 proteins with the right topology, (Table 2). For the same dataset, TMBpro identifies the correct topology for 19 proteins. Thus, clearly tobmodel has a higher accuracy in terms of topology predictions. Moreover, for the dataset of 23 proteins used for the development of 3D-SPoT (Naveed *et al.*, 2012), tobmodel predicts the correct topology for 18 proteins.

3.2 Z-coordinate prediction using ZPRED3

ZPRED3 is a Z-coordinate predictor implemented for estimating the distance of a residue from the membrane center. Different input features were tried as inputs to the SVMs. Furthermore, we also tried different SVM parameters and window sizes. As shown in Table 1, a combination of a PSSM and predicted topology obtained from BOCTOPUS provide the lowest average error, 1.61 Å, and the highest correlation coefficient, 0.78. Further, approximately 74% residues were correctly predicted to be within 2 Å of their observed Z-coordinate. Figure 2 shows the ZPRED3 and BOCTOPUS output for the FadL outer membrane protein from *Pseudomonas aeruginosa*. ZPRED3 predictions are shown with dashed lines and the real Z-coordinate in a continuous line (Fig. 2). As can be seen, the correlation in the pre-barrel stage is low, but within the barrel it is quite accurate. This indicates that a separate pre-barrel region identifier might be needed to generate accurate predictions for the entire protein.

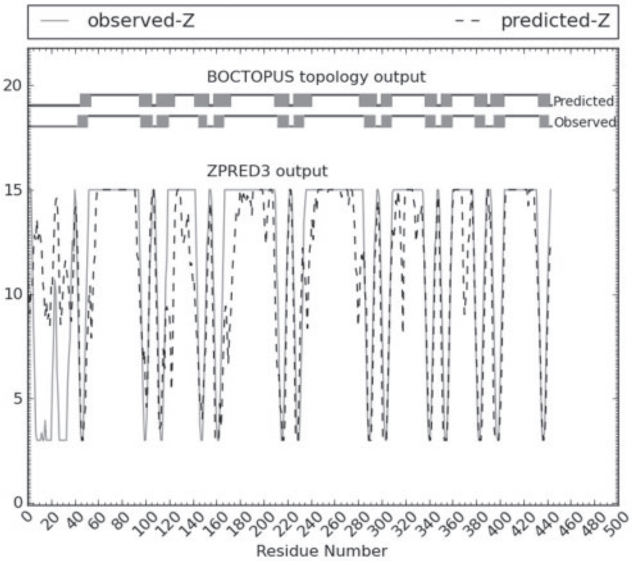

Fig. 2. BOCTOPUS and ZPRED3 output for the FadL outer membrane protein (FadL) from *Pseudomonas aeruginosa* (3DWO_X). The *X*-axis shows the residue number. Observed and predicted Z-coordinates are shown in dark and gray lines. The most likely BOCTOPUS topology prediction is shown with horizontal bars. Outer loops, inner loops and the TM strands are shown in gray color at different heights.

Figure 3 shows the average Z-coordinate error for all 540 transmembrane β-strands in our dataset of 36 proteins (Hayat and Elofsson, 2012). The data are divided into two subsets; the correctly and incorrectly identified β-strands. The Z-coordinate error for correct strands is significantly lower (1.7 Å) than for the incorrectly predicted strands (9.1 Å).

When ranking the final models, we used the average difference between the Z-coordinate predicted by ZPRED3 and the Z-coordinate in a model. Here, we have simply identified the top ranking model to be the model with the lowest average difference between the Z-coordinates. The $C\alpha$ models, generated as described above, were ranked based on the minimum error between the predicted Z-coordinate obtained from ZPRED3 and the Z-coordinate obtained from the $C\alpha$ model, i.e. the Z-coordinate error.

3.3 Comparison with TMBpro and 3D-SPoT

Here, we present a comparison of the models generated by tobmodel, TMBpro (Randall *et al.*, 2008) and 3D-SPoT (Naveed *et al.*, 2012). Only tobmodel and TMBpro predict topologies. The accuracy is clearly higher for tobmodel, 27 versus 19 correct predictions, (Table 2). Also, the number of proteins with correct topologies selected by tobmodel is slightly higher than for BOCTOPUS

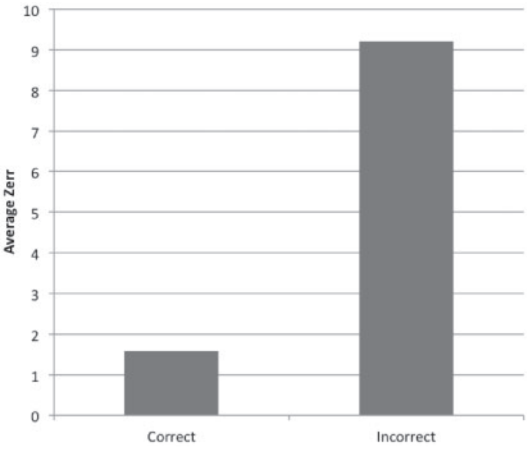

Fig. 3. Average Z-coordinate prediction errors for BOCTOPUS predictions. Comparison of average errors in Z-coordinate prediction based on correctness of the location of the identified transmembrane β-strand. Incorrect strands refers to under predicted, over predicted and strands that are predicted at a location that does not overlap with their observed location in the structure

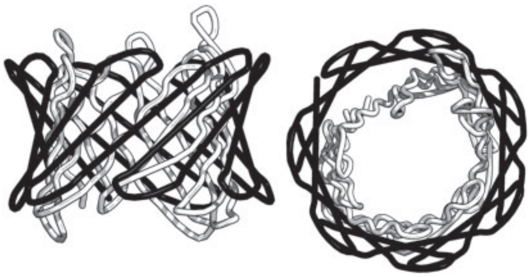

Fig. 4. The top ranking model of the FadL outer membrane protein (FadL) from *Pseudomonas aeruginosa* (3DWO_X) as generated by tobmodel is shown in black and the actual known structure obtained from the OPM database is shown in white. Figure on the left and right are the front and the top view, respectively. Only the barrel part is shown. The $C\alpha$ RMSD is 7.56 Å.

Table 3. Topology comparison of the top ranking $C\alpha$ model

Protein (obs. Strands)	tobmodel (Top ranking)			TMBpro (web server)		
	Pred. strands	RMSD	TM_Score	Pred. strands	RMSD	TM_Score
1qj8_A (8)	8	8.13	0.26	8	4.85	0.51
1e54_E (16)	12	10.20	0.28	16	8.74	0.78
1a0s_R (18)	18	8.72	0.39	18	6.13	0.62
1i78_A (10)	10	7.91	0.29	12	13.14	0.41
2qdz_A (16)	16	7.95	0.50	22	22.96	0.30
2wjr_A (12)	12	5.89	0.37	12	5.56	0.60
2j1n_C (16)	16	7.42	0.40	16	5.37	0.65
2ysu_A (22)	22	3.29	0.76	12	12.15	0.51
1k24_A (10)	10	7.26	0.31	10	4.59	0.71
1p4t_A (8)	8	7.67	0.28	8	3.80	0.69
3kvn_A (12)	12	9.06	0.25	18	23.95	0.34
2k0l_A (8)	8	4.93	0.44	10	13.29	0.41
2iah_A (22)	20	4.98	0.67	22	7.27	0.65
3a2s_G (16)	16	9.73	0.38	16	5.28	0.61
2f1v_A (8)	8	5.23	0.44	8	3.24	0.74
3csl_A (22)	20	10.65	0.53	22	7.83	0.60
3fhh_A (22)	22	3.69	0.72	22	7.12	0.68
2o4v_C (16)	16	10.30	0.40	18	9.68	0.42
2vqi_A (24)	20	8.99	0.42	22	16.87	0.26
3dwo_X (14)	14	7.56	0.35	16	9.95	0.45
1qd6_D (12)	12	7.29	0.33	12	4.66	0.73
1tly_A (12)	12	8.30	0.41	12	5.47	0.64
2mpr_C (18)	18	8.36	0.51	18	3.64	0.77
3prn_C (16)	14	7.49	0.33	18	9.95	0.41
3dzm_A (8)	8	9.45	0.26	8	16.97	0.32
1fep_A (22)	22	3.43	0.75	22	5.44	0.79
2erv_A (8)	8	10.93	0.21	8	7.54	0.47
3bs0_A (14)	14	6.33	0.41	14	4.77	0.65
2por_C (16)	16	6.44	0.42	16	4.39	0.62
1t16_A (14)	14	7.88	0.36	16	9.90	0.45
2qom_A (12)	12	8.69	0.29	12	7.52	0.41
1uyo_X (12)	12	5.52	0.46	14	12.52	0.39
1kmp_A (22)	20	4.87	0.67	22	10.48	0.62
2iww_A (14)	14	4.41	0.59	14	4.31	0.69
3jty_A (18)	18	5.55	0.61	16	9.94	0.44
2grx_A (22)	20	6.09	0.56	22	7.18	0.76

Analysis of top ranking models generated by tobmodel chosen from 10000 topologies obtained from BOCTOPUS. Tobmodel and TMBpro predict the correct number of β-strands for 29 and 24 proteins in the dataset, respectively.

[a] Here, in the case of TMBpro, Z-coordinate error refers to the average error between the actual Z-coordinate obtained from the known structure and the model obtained from TMBpro web server. For tobmodel, the Z-coordinate error is the same as defined in Section 2. Tobmodel results are based on a strict 10-fold cross-validation both during BOCTOPUS and ZPRED3 stages. RMSD and TM_Score of the top ranking tobmodel models and models obtained from TMBpro web server are calculated with respect to the known structure obtained from the OPM database. TM_Score is calculated using the TM-score method (Zhang and Skolnick, 2004). For the known structure, only the $C\alpha$ atoms predicted to be in the TM region were taken into consideration for calculating the TM_Score. The average TM_Score for tobmodel and TMBpro is 0.43 and 0.56, respectively.

(Hayat and Elofsson, 2012). Further, to access the quality of the generated models, the average RMSD and TM_Score (Zhang and Skolnick, 2004) of the top ranking models are reported in Table 3. The average TM_Score for tobmodel and TMBpro is 0.43 and 0.56, respectively, while the average RMSD is 7.24 and 8.79 Å. For the 27 proteins for which tobmodel selects the correct topology, the average RMSD drops to 6.94 Å. Figure 4 shows the top ranking model obtained for the FadL outer membrane protein (FadL) from Pseudomonas aeruginosa (3DWO$_X$). The RMSD distribution of all models generated and the top ranking models selected by tobmodel is shown in Figure 5.

Since the models generated by 3D-SPoT were not available, only the RMSD reported in the article on the 3D-SPoT dataset (Naveed *et al.*, 2012) could be used for comparison. In addition, 3D-SPoT does not predict topologies, but rather uses an available correct topology. Thus, results for 3D-SPoT are reported for the case when all topologies are correct. Therefore, for a fair comparison, we have used both predicted and correct topologies to make the models using tobmodel, (Table 3). The average RMSD of tobmodel on this dataset is 7.06 Å, which is significantly less accurate than the reported values for 3D-SPoT of 4.10 Å. However, when the correct topologies were used by tobmodel, the RMSD of top ranking models drops to 5.86 Å. It would appear that 3D-SPoT generates more accurate models than tobmodel. This is likely due to that tobmodel currently does not take elliptical shaped barrels into account (Figure 6).

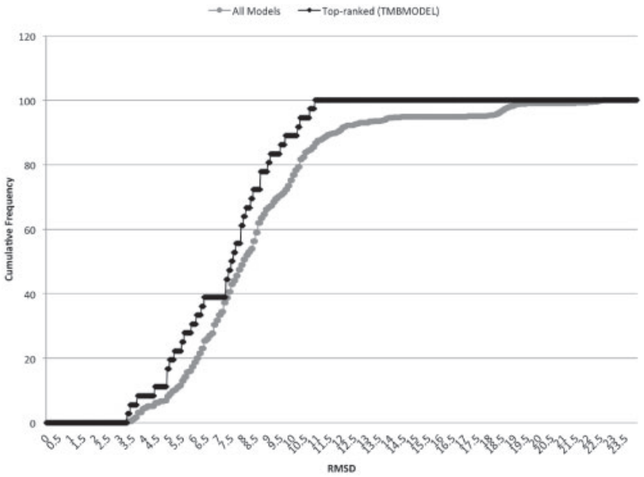

Fig. 5. The RMSD distribution of top ranking models is slightly better than the average for all models. Almost 40% of the top ranking models have an RMSD of ≤ 6.5 Å, while the corresponding number for all models generated by tobmodel is 21%

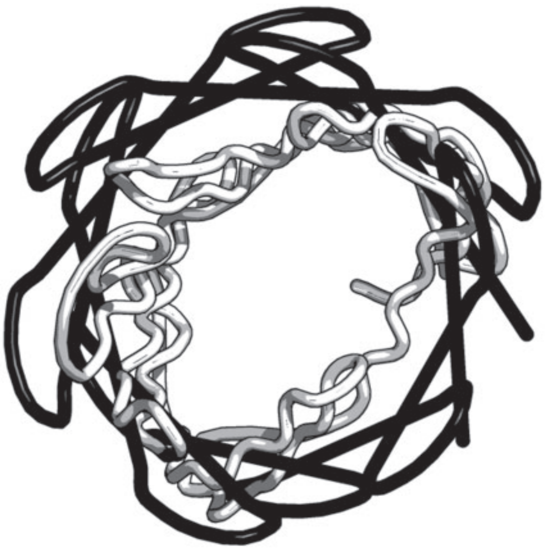

Fig. 6. The top ranking tobmodel model for outer membrane phospholipase A in *Escherichia coli* (PDB ID: 1QD6_D) is shown in black and the known structure is shown in white in top view. As can be seen here, tobmodel does not accurately generate models of proteins that have an elliptical shape

3.4 Modeling VDAC

Finally, when modeling VDAC1, the only known Transmembrane β-barrel with an odd number of strands, tobmodel always predicts it to have 20 instead of 19 strands, resulting in a model with 16.47 Å RMSD and 0.21 TM_score. However, if we use the correct topology, tobmodel generates a model with RMSD of 2.86 Å and a TM_score 0.75.

4 CONCLUSION

Here, we present a novel method, tobmodel, that given the amino acid sequence for a β-barrel membrane protein generates a $C\alpha$ model. First, hundreds of alternative topologies are produced by means of our β-barrel topology predictor, BOCTOPUS. Subsequently, many alternative $C\alpha$ models are generated for each topology. Finally, the best of all these models is selected based on the agreement between a model and the predicted distance (Z-coordinate) from the center of the membrane. This distance is predicted using a novel support vector machine-based predictor, ZPRED3. The selection procedure results in a slightly higher number of correct topologies compared to BOCTOPUS alone.

Tobmodel performs well compared to available methods. Although the quality of the models generated by tobmodel is somewhat worse than models generated by 3D-SPoT (Naveed *et al.*, 2012), it is on par with models by TMBpro (Randall *et al.*, 2008), the only previously published method for β barrel membrane protein prediction that takes protein sequences as input rather than pre-computed topologies. Further, the average RMSD of top ranking models, generated by tobmodel, is 7.24 Å. Taken together, our results indicate that the inclusion of Z-coordinate errors in the modeling of transmembrane β-barrels is a promising new direction and, further, suggest that additional selection criteria and more accurate modeling may further improve our method.

ACKNOWLEDGEMENT

We are grateful for Dr Sara Light who helped with proof reading the paper. SH is thankful to K & A Wallenbergs Stipend to attend ISMB 2012.

Funding: This work was supported by grants from Swedish Research Council [VR-NT 2009-5072, VR-M 2010-3555], SSF (the Foundation for Strategic Research) and Vinnova through the Vinnova-JSP program, the EU 7'Th Framework Program by support to the EDICT project, contract No: FP7-HEALTH-F4-2007-201924.

Conflict of Interest: none declared.

REFERENCES

Altschul,S. *et al*. (1997) Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Bagos,P. *et al*. (2004) PRED-TMBB: a web server for predicting the topology of β-barrel outer membrane proteins. *Nucleic Acids Res.*, **32**, W400.

Bagos,P. *et al*. (2005) Evaluation of methods for predicting the topology of β-barrel outer membrane proteins and a consensus prediction method. *BMC Bioinformatics*, **6**, 0–7.

Becker,T. *et al*. (2012) Mitochondrial protein import: from transport pathways to an integrated network. *Trends Biochem. Sci.*, **37**, 85–91.

Bigelow,H. and Rost,B. (2006) PROFtmb: a web server for predicting bacterial transmembrane beta barrel proteins. *Nucleic Acids Res.*, **34**(Suppl. 2), W186.

Chou,K. *et al*. (1984) Energetic approach to the packing of. alpha.-helixes. 2. general treatment of nonequivalent and nonregular helixes. *J. Am. Chem. Soc.*, **106**, 3161–3170.

Chou,K. *et al*. (1990) Conformational and geometrical properties of idealized β-barrels in proteins. *J. Mol. Biol.*, **213**, 315–326.

Freeman,T. and Wimley,W. (2010) A highly accurate statistical approach for the prediction of transmembrane β-barrels. *Bioinformatics*, **26**, 1965.

Galdiero,S. *et al*. (2007) β-barrel membrane bacterial proteins: structure, function, assembly and interaction with lipids. *Curr. Protein Peptide Sci.*, **8**, 63–82.

Granseth,E. *et al*. (2006) Zpred: predicting the distance to the membrane center for residues in α-helical membrane proteins. *Bioinformatics*, **22**, e191–e196.

Gromiha,M. *et al*. (2004) Neural network-based prediction of transmembrane β-strand segments in outer membrane proteins. *J. Comput. Chem.*, **25**, 762–767.

Gromiha,M. *et al*. (2005) TMBETA-NET: discrimination and prediction of membrane spanning β-strands in outer membrane proteins. *Nucleic Acids Res.*, **33**(Suppl. 2), W164.

Gront,D. *et al*. (2007) Backbone building from quadrilaterals: A fast and accurate algorithm for protein backbone reconstruction from alpha carbon coordinates. *J. Comput. Chem.*, **28**, 1593–1597.

Hayat,S. and Elofsson,A. (2012) BOCTOPUS: improved topology prediction of transmembrane β barrel protein. *Bioinformatics*, **28**, 516–522.

Koebnik,R. *et al*. (2000) Structure and function of bacterial outer membrane proteins: barrels in a nutshell. *Mol. Microbiol.*, **37**, 239–253.

Lomize,M. *et al*. (2006) OPM: orientations of proteins in membranes database. *Bioinformatics*, **22**, 623–625.

Martelli,P. *et al*. (2002) A sequence-profile-based HMM for predicting and discriminating β barrel membrane proteins. *Bioinformatics*, **18**(Suppl. 1), S46.

Mirus,O. and Schleiff,E. (2005) Prediction of-barrel membrane proteins by searching for restricted domains. *BMC Bioinformatics*, **6**, 254.

Murzin,A. *et al*. (1994a) Principles determining the structure of [beta]-sheet barrels in proteins i. a theoretical analysis. *J. Mol. Biol.*, **236**, 1369–1381.

Murzin,A. *et al*. (1994b) Principles determining the structure of [beta]-sheet barrels in proteins ii. the observed structures. *J. Mol. Biol.*, **236**, 1382–1400.

Naveed,H. *et al*. (2012) Predicting three-dimensional structures of transmembrane domains of -barrel membrane proteins. *J. Am. Chem. Soc.*, **134**, 1775–1781.

Pajón,R. *et al*. (2006) Computational identification of beta-barrel outer-membrane proteins in Mycobacterium tuberculosis predicted proteomes as putative vaccine candidates. *Tuberculosis*, **86**, 290–302.

Papaloukas,C. *et al*. (2008) Estimating the length of transmembrane helices using z-coordinate predictions. *Protein Sci.*, **17**, 271–278.

Randall,A. *et al*. (2008) TMBpro: secondary structure, β-contact and tertiary structure prediction of transmembrane β-barrel proteins. *Bioinformatics*, **24**, 513–520.

Remmert,M. *et al*. (2009) HHomp - prediction and classification of outer membrane proteins. *Nucleic Acids Res.*, **37**(Suppl. 2), W446.

Schulz,G. (2002) The structure of bacterial outer membrane proteins. *BBA-Biomembranes*, **1565**, 308–317.

Singh,N. *et al*. (2011) TMBHMM: a frequency profile based HMM for predicting the topology of transmembrane beta barrel proteins and the exposure status of transmembrane residues. *Biochim. Biophys. Acta.*, **1814**, 664–670.

Waldispühl,J. *et al*. (2008) Modeling ensembles of transmembrane β-barrel proteins. *Proteins*, **71**, 1097–1112.

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

Yan,R. *et al*. (2011) Outer membrane proteins can be simply identified using secondary structure element alignment. *BMC Bioinformatics*, **12**, 76.

Zhang,Y. and Skolnick,J. (2004) Scoring function for automated assessment of protein structure template quality. *Proteins*, **57**, 702–710.