BIOINFORMATICS ORIGINAL PAPER Vol. 28 no. 4 2012, pages 516–522
doi:10.1093/bioinformatics/btr710

Structural bioinformatics Advance Access publication January 13, 2012

BOCTOPUS: improved topology prediction of transmembrane β barrel proteins

Sikander Hayat and Arne Elofsson*

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

Associate Editor: Burkhard Rost

ABSTRACT

Motivation: Transmembrane β barrel proteins (TMBs) are found in the outer membrane of Gram-negative bacteria, chloroplast and mitochondria. They play a major role in the translocation machinery, pore formation, membrane anchoring and ion exchange. TMBs are also promising targets for antimicrobial drugs and vaccines. Given the difficulty in membrane protein structure determination, computational methods to identify TMBs and predict the topology of TMBs are important.

Results: Here, we present BOCTOPUS; an improved method for the topology prediction of TMBs by employing a combination of support vector machines (SVMs) and Hidden Markov Models (HMMs). The SVMs and HMMs account for local and global residue preferences, respectively. Based on a 10-fold cross-validation test, BOCTOPUS performs better than all existing methods, reaching a Q3 accuracy of 87%. Further, BOCTOPUS predicted the correct number of strands for 83% proteins in the dataset. BOCTOPUS might also help in reliable identification of TMBs by using it as an additional filter to methods specialized in this task.

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

Contact: arne@bioinfo.se

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

Received on July 6, 2011; revised on December 14, 2011; accepted on December 20, 2011

1 INTRODUCTION

Two types of transmembrane proteins are known to exist, α-helical membrane proteins and transmembrane β barrel proteins (TMBs). The former class has been the major focus of computational methods, as they are known to constitute 20–30% of a typical proteome (Wallin and von Heijne, 1998). However, TMBs are worthy of attention as they play an important role in the translocation machinery of both inner and outer membrane proteins in bacteria, chloroplast and mitochondria. Moreover, TMBs are involved in transport of molecules, voltage gating, membrane anchoring, pore formation and are also candidate molecular targets for development of antimicrobial drugs and vaccines (Galdiero *et al.*, 2007; Koebnik *et al.*, 2000; Pajón *et al.*, 2006; Schulz, 2002). However, the number of solved TMB structures in PDB is limited, since they are difficult to crystalize; consequently, better computational methods for the topology prediction of TMBs are needed. Such a predicted topology can function as a template for experimental investigations and can further aid in elucidating the structure and function of putative TMBs.

TMBs typically consist of a central pore region made up of anti-parallel β strands and residues in the β strands follow a strict dyad repeat pattern (Seshadri *et al.*, 1998). The general construction principles of TMBs (Schulz, 2002) have been employed for the topology prediction and genome-wide identification of TMBs. However, the low number of known TMB structures and a less prominent hydrophobicity profile pose problems in the development of computational methods for the identification and topology prediction of TMBs.

The computational methods in the realm of TMBs can be divided into two parts (i) methods that aim to identify TMBs from genomic data and (ii) methods that predict the TMB topology assuming that the given sequence is a putative TMB. The first group consists of a variety of methods including methods that combine statistical propensities and C-terminal pattern identification (Berven *et al.*, 2004), empirical scores (Freeman and Wimley, 2010; Mirus and Schleiff, 2005; Wimley, 2002), *K*-nearest neighbor methods (Hu and Yan, 2008), SVMs (Park *et al.*, 2005), Neural Networks (Gromiha and Suwa, 2006; Gromiha *et al.*, 2004), Hidden Markov Models (HMM) (Deng *et al.*, 2004; Martelli *et al.*, 2002), amino acid composition (Garrow *et al.*, 2005; Gromiha *et al.*, 2005a) and secondary structure element alignments (Yan *et al.*, 2011). The HHomp method for the identification of TMBs employs HMM-profile comparison and is based on the observation that almost all β-barrel OMP have a common ancestry (Remmert *et al.*, 2009).

Methods aiming at the prediction of topologies include HMM-based methods such as PRED-TMBB (Bagos *et al.*, 2004), TMB-HMM (Singh *et al.*, 2011) and PROFtmb (Bigelow and Rost, 2006), SVM-based methods such as TMBETAPRED-RBF (Ou *et al.*, 2010), neural network-based methods such as TMBpro (Randall *et al.*, 2008) and methods based on statistical potentials such as transFold (Waldispuhl *et al.*, 2006). PROFtmb (Bigelow and Rost, 2006), TMBETA-NET (Gromiha *et al.*, 2005b) and PRED-TMBB (Bagos *et al.*, 2004) also predict the topology of the identified TMBs. A comparison and evaluation carried out indicated that HMM-based methods outperform methods based on other types of machine learning (Bagos *et al.*, 2005).

Here, we present an improved topology predictor for TMBs named BOCTOPUS. BOCTOPUS is based on the ideas used in two

*To whom correspondence should be addressed.

recently developed methods for the topology prediction of α-helical membrane proteins; MEMSAT-3 (Jones, 2007) and OCTOPUS (Viklund and Elofsson, 2008). BOCTOPUS uses a combination of SVMs to predict the local structural preferences for a residue, and a HMM model to create a topology model for a protein. Based on a 10-fold cross-validation test, BOCTOPUS predicted the correct number of transmembrane β strands in 30 out of 36 TMBs in the dataset and achieved an overall Q3 accuracy of 87%. The primary use of BOCTOPUS is topology prediction of TMBs with the assumption that all input sequences are TMBs. However, we show that BOCTOPUS can reduce the number of false positives when it is used along with specialized methods for TMB identification such as BOMP (Berven *et al.*, 2004) or PSORTb (Yu *et al.*, 2010).

2 METHODS

2.1 Training dataset

A non-redundant dataset was obtained from the OPM database (Lomize *et al.*, 2006). The dataset was homology reduced at $\leq$30% sequence identity resulting in 36 TMB structures. Development and training of the SVMs and HMMs used in BOCTOPUS was performed based on 10-fold cross-validation. To further avoid influence by distantly related homologs, the training was performed in such a way that all proteins belonging to the same OPM family were put together in the same cross-validation set.

All residues in the dataset were annotated as either 'I' (inner-loop), 'O' (outer-loop) or 'M' (transmembrane β strand) based on the coordinate of the $C\alpha$ atoms and membrane boundaries obtained from the OPM database (Lomize *et al.*, 2006). Here, residues located within the membrane boundaries but do not belong to a transmembrane β-strand are labeled as 'I' or 'O' based on the location of the initial residue. The annotated dataset is available from the web server.

2.2 Training BOCTOPUS

The architecture of BOCTOPUS consists of two layers (Fig. 1). The first layer consists of three SVMs that predict the local preferences for a residue to be in a particular location. The second layer consists of an HMM model that predicts the topology. The dataset was divided into 10 sets such that proteins belonging to the same super-family were in the same set. During the training, nine sets were used to test the performance on the 10th set. In contrast to what has been done in most earlier studies, this cross-validation was maintained throughout the entire pipeline.

2.2.1 Input features The input feature for BOCTOPUS is a position-specific scoring matrices (PSSMs) obtained using PSI-BLAST version 2.2.18 (Altschul *et al.*, 1997). Here, default parameters and three iterations of searching the non-redundant nr-database, obtained from the NCBI website in July 2010, was used. The log-odds value in the PSSM was transformed into a PSSM profile by dividing all number by 10 such that they lie between $\pm$1.0.

2.2.2 SVMs training Three SVMs, as implemented in the libsvm interface in the R e1071 package were trained to determine the preference of each residue to be in the 'I', 'O' or "M" regions. Radial basis and linear kernels, different windows sizes in the range of 1– 31 were tried (Supplementary Material). The optimal window size was determined based on the highest Matthews correlation coefficient (MCC) (Matthews, 1975).

2.2.3 Optimization of the HMMs 'IOM-profile' generated from the probabilities produced by the three SVMs was used as the input for training different combinations of HMM parameters. HMMs used in BOCTOPUS

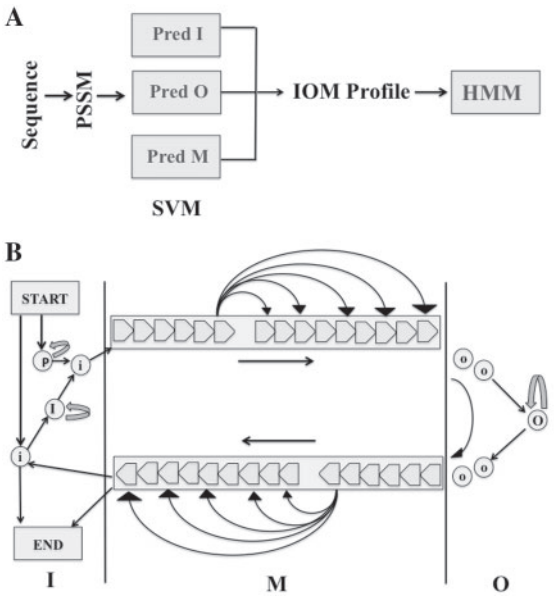

Fig. 1. BOCTOPUS pipeline. Psi-blast is used to generate PSSM for a given sequence. A) Three separate SVMs are used to predict the residue-level preference for each amino acid to be I, M and O, respectively. An 'IOM-profile' is generated from the probabilities obtained from the SVMs. B) The 'IOM-profile' is then used by an HMM to predict the global topology. The HMM architecture is explained in Section 2.2.3. The final topology is calculated using the Viterbi algorithm 2.2.3.

are implemented in the modhmm package (Viklund and Elofsson, 2004) and the HMM architecture is shown in Figure 1. The HMM describing the global topology consists of a pre-barrel stage (P) describing the region before the first transmembrane β-strand is detected. Further, a TMB is defined by four different states each representing the inner-loop, outer loop and the up and the down strands (Fig. 1). The up and down strand states can handle β-strands in the range of 6–15 residues. To be consistent with structural properties known from the available 3D structures, all protein topologies start in the 'P' or 'i' state and end in the 'M' (down strand) or 'i' state. The emission scores for the states are the probabilities obtained from the respective SVMs. Based on the emission scores, the most likely topology is predicted using the Viterbi algorithm.

The transition probabilities between states are set to 1.0 to make the final predicted topology dependent only on the SVM output values and the HMM architecture, but not on the distribution of topologies in the training dataset. Thereby fewer parameters need to be optimized. The emission scores for the 'I', 'O' and 'M' states are directly set to the probability scores obtained from the respective SVMs. However, to accommodate for the variable length of large outer-loops, small inner-loops and pre-barrel (defined as the region before the first transmembrane beta-strand) part of the sequence, we found that it was necessary to optimize the three states with self-loops. These three states are shown in bold letters I, P and O with a self-loop (Fig. 1). The four parameters tested per-state are the weights for 'I' and 'O' emissions in determining the emission score and the transition probability to go to the next state or self-loop. Each parameter was tested for values in the range 0–1 with a step size of 0.1. Initially, the four parameters for each state (i.e. I, P, O) were optimized separately, keeping the parameters of the other two states fixed at 0.5 using the same cross-validation scheme as when optimizing the SVMs. The best performing HMM parameters were chosen based on the correct number of predicted strands on the training set. The best HMM parameterspobtained for I and O states were then combined and IO-optimized HMMs were obtained. IO-optimized HMMs were then

combined in all possible combinations with parameters obtained for the P state.

2.3 Global topology prediction based on cross-validated models

As mentioned above, for each round, training was performed on nine sets, and the remaining set was used for testing. First 'IOM-profiles' were generated for proteins in the test-set using SVMs trained only on the training sets. However, as thousands of parameters provided identical (and perfect) results on the training set, the topologies of the proteins in the test-set were then determined by using a subset of these top performing HMMs (for details see Supplementary Material). For the final evaluation, 10 000 HMMs were randomly selected from a pool of the top performing HMMs to predict the topology of the proteins in the test sets.

2.4 Evaluation

For per-residue accuracy performance Q2, Q3 and segment overlap (SOV) (Rost *et al.*, 1994) were used. Q2 is defined as the two-state (membrane/not-membrane) prediction accuracy. Q3 is defined as the three-state prediction accuracy for i, M, o states. In addition, the number and location of the predicted strands was used to evaluate the performance per-protein. A protein was defined to have a correct predicted topology when the number of predicted strands is correct and each predicted strand overlaps with at least two residues with the observed strand. It should be noted that all results for BOCTOPUS are based on the 10-fold cross-validation and are reported as an average of a randomly selected set of 10 000 among the top performing HMM parameters.

2.5 TMB identification

The ability to identify TMBs was tested on a non-redundant (at sequence identity ≤50%) representative dataset of 14 232 PDB entries obtained from Freeman and Wimley (2010). Here, a protein was assigned as a TMB when the number of strands predicted by BOCTOPUS is larger than a given number (typically 8). Methods such as BOMP (Berven *et al.*, 2004), PSORTb (Yu *et al.*, 2010) and tmbetaNet (Gromiha *et al.*, 2005b) have also been tested in combination with BOCTOPUS for the identification of putative TMBs.

3 RESULTS AND DISCUSSION

3.1 Residue-level prediction accuracy

BOCTOPUS is a two stage computational method for the topology prediction of TMBs. In the first stage, three different SVMs are trained to distinguish between β-strand/not β-strand, inner-loop/not inner-loop, outer-loop/not outer-loop regions. Different window sizes ranging from 1 to 31 were tested (Fig. 2).

The optimal window size was determined based on the MCC values such that no statistically significant improvement was gained on further increasing the window size. Based on this criteria, window size of 31, 19 and 21 was chosen for i, o and M SVMs, respectively. *P*-values for determining the statistical significance of change in MCC values with increasing window size are given in the supplementary information. This window size was maintained in all cross-validation sets; however, the SVMs were trained separately in each training set.

For each residue, the probabilities obtained from the individual SVMs are used to generate an 'IOM-profile', which are fed into the HMM stage to produce the final prediction.

When calculating the accuracy measures using the SVMs alone, each residue is assigned to the region with the highest probability

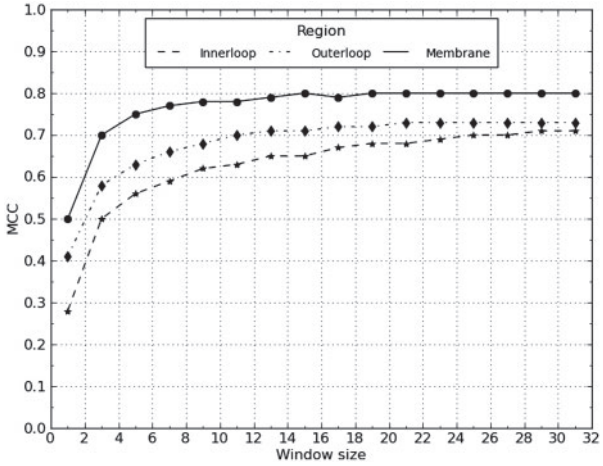

Fig. 2. BOCTOPUS per-residue preference prediction. MCC values at different window sizes in the range of 1–31 are shown for per-residue i, M, o prediction. Window size of 31, 21 and 19 is chosen to generate residue preference based on separate SVMs for i, M and o regions, respectively (Section 3.1). The resulting probabilities for each state preference are combined into an 'IOM-profile' and then used by an HMM to predict the global topology.

Table 1. Per residue accuracy

Method	TP	FP	FN	TN	Q2 (%)	Q3 (%)	SOV (%)
BOCTOPUS	4471	634	538	8335	91	87	92
BOCTOPUS-SVM	442	663	690	8183	90	85	75
PRED-TMBB	4288	817	1236	7637	85	82	85
PROFtmb[a]	4832	273	1922	6951	84	–	90
TMBETAPRED-RBF[a]	4019	966	899	7829	85	–	87
TMBpro	4019	645	1448	7425	85	81	86

Per-residue accuracy comparison. Q2, Q3 and SOV scores for BOCTOPUS and other methods. BOCTOPUS-SVM shows the accuracy measures before the HMM stage. True positive (TP), false positive (FP), false negative (FN), true negative (TN) are reported for Q2. All BOCTOPUS results are reported based on the 10-fold cross-validation test. [a]The output of these methods is only two-state (membrane/not-membrane), i.e. Q3 can not be calculated.

in the 'IOM-profile'. The Q2, Q3 and SOV scores for BOCTOPUS-SVM are 90, 85 and 75%, respectively (Table 1). The Q2 and Q3 scores compete favorably with earlier methods, which have Q2 scores ∼85% and Q3 scores up to 82%. However, the SOV score is much lower than these methods, as most of the strands predicted are too short.

3.2 Topology prediction using BOCTOPUS

In BOCTOPUS, the SVM predictions are used as input into a HMM-like model to obtain the final prediction. This step increases the per-residue accuracy in particular as measured by Q3 and particularly SOV (Table 1). It should be noted that, as many parameters performed equally well in the optimization of the HMMs, the predictions of the 10 test-sets shown here are the average performance of a randomly selected set of 10 000 among the top performing HMM parameters. The SOV score (92%) of BOCTOPUS is higher than any of the earlier methods. BOCTOPUS

Table 2. Topology prediction comparison

Methods	No. strands	Topology	UP	OP	FS (%)
BOCTOPUS	30.1 ± 1.5	25.4 ± 2.0	3	2	96
PRED-TMBB	22	15	5	9	86
PROFtmb	27	25	4	5	76
TMBETAPRED-RBF[a]	27	21	4	4	76
TMBpro	24	19	3	9	76

Comparison of topology predictions. No. strands is defined as the number of sequences where the number of predicted strands is equal to the number of observed strands. Topology is defined as if number of strands is correct and the predicted strands overlap by at least two residues. Underpredicted (UP) is defined as sequences where the number of strands underpredicted, i.e. some strands are missed. Overpredicted (OP) is defined as sequences where the number of strands is overpredicted. BOCTOPUS-SVM shows the accuracy measures without the HMM stage. Fraction of strands (FS) is defined as the number of observed strands that are correctly predicted to be at the correction location. The total number of strands in the dataset is 540. BOCTOPUS results are reported as an average of a randomly selected set of 10 000 among the top performing HMM parameters and are based on the 10-fold cross-validation test. Detailed 10-fold cross-validation test results per protein are given in Table 3.
[a]TMBETAPRED-RBF (Ou *et al.*, 2010) classified 2qomA as a non-TMB protein.

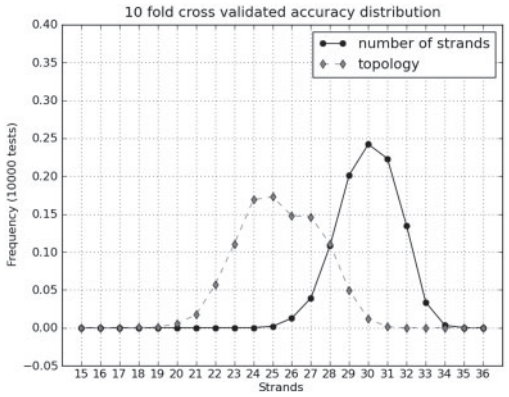

Fig. 3. The accuracy distribution of BOCTOPUS HMMs on test-sets. On average, BOCTOPUS predicts the correct number of strands and correct topology for 30.1 ± 1.5 and 25.4 ± 2 proteins out of 36 proteins, respectively.

predicts the correct number of strands on average for 30.1 out of 36 proteins (Table 2). Further, BOCTOPUS predicted on average 25.4 of these proteins with correct topology. Here, a topology is defined as correct when the number of predicted strands is equal to the number of observed strands and all predicted strands overlap the observed strand. Figure 3 shows the topology and correct number of strands distribution for BOCTOPUS run on a randomly selected set of 10 000 among the top performing HMM parameters. As shown, almost 65% of the 10 000 HMMs tested can predict the correct number of strands for >30 out of 36 proteins (Fig. 3).

In Figure 4, the most frequent incorrect predictions are shown. The top six are proteins for which the number of strands was predicted correctly but at least one predicted strand does not overlap with its observed location. In 9 of 12 cases, the errors can be attributed to overpredictions in the pre-barrel state (Table 3); whereas, for 1a0s_A and 2mpr_C, one long outer-loop is predicted

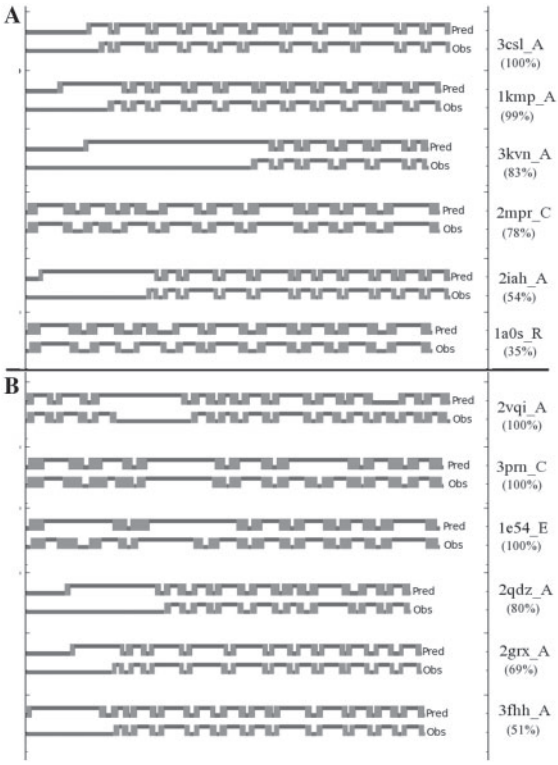

Fig. 4. Erroneous topology predictions using BOCTOPUS. (**A**) In the top six cases, BOCTOPUS predicts the correct number of strands; however, one or more strands does not overlap with observed strands. (**B**) In the bottom six cases, the number of strands predicted by BOCTOPUS does not match the number of observed strands. The number below the pdb id represents the percentage of incorrect outcomes when multiple HMMs were employed from the pool of best-trained HMMs. All results reported are based on 10-fold cross-validation test. Only the most frequently occurring error is shown per protein. Topology errors per protein and the most often occurring topology errors are given in the Supplementary Material.

shorter than observed, resulting in one predicted strand to be shifted. BOCTOPUS underpredicts the number of strands in two proteins (2vqi_A and 1e54_E). These two proteins are also missed by PROFtmb and TMBETAPRED-RBF that show the second highest accuracy on our dataset (Table 2). For 2vqi_A, BOCTOPUS misses two strands at the C-termini, in 3prn_C, Strands 10 and 11 are always missed and for 1e54_E only 12 of 16 observed strands are predicted correctly.

3.3 Multi-chain TMB topology prediction

Multi-chain TMBs are TMBs whose barrel composes of β-strands from different chains. Such multi-chain TMBs were not discussed in Section 3.2 because the grammar of multi-chain TMBs differs from that of single-chain TMBs (Bigelow and Rost, 2006; Remmert *et al.*, 2010). For example, TOLC protein from *Escherichia coli* (1tqq_A), Drug-Discharge Outer Membrane Protein, OprM from *Pseudomonas aeruginosa* (1wp1_A), Multidrug Resistance (VceC) protein from *Vibrio cholerae* (1yc9_A) have very long inner-loops. A comparison of different prediction methods on multi-chain TMBs shows that none of the methods, including BOCTOPUS, can predict

Table 3. Per protein topology prediction accuracy

PDB ID	Topology	No. strands	UP	OP	OV	Fraction (%)
3prn_C	0.0	0.0	10/11	–	–	87
1e54_E	0.0	12.14	2/3/6/7		–	52
2vqi_A	0.0	14.33	19/20	–	6	45
3csl_A	0.0	73.03	–	–	1/11	76
1kmp_A	1.02	83.92	–	–	1	89
3kvn_A	16.28	70.37	–	–	1	64
2qdz_A	19.6	19.6	–	1/2	–	100
2mpr_C	20.53	100.0	–	–	6	100
2grx_A	30.58	65.02	–	–	1	53
2iah_A	45.39	96.5	–	–	1	97
3fhh_A	48.32	48.32	–	1/2	–	100
1a0s_R	63.96	100.0	–	–	6	100
3bs0_A	70.28	70.28	–	–	1/2	100
1fep_A	71.81	71.81	–	–	1/2	100
3a2s_G	79.46	95.9	–	–	5	59
2por_C	90.09	90.09	–	–	6	100
1t16_A	93.03	93.03	–	–	1/2	100
3dzm_A	94.01	100.0	–	–	3	100
3dwo_X	97.22	97.22	–	–	1/2	100
2o4v_C	97.31	99.47	–	–	12	86
2ysu_A	100.0	100.0	–	–	–	–
2wjr_A	100.0	100.0	–	–	–	–
2erv_A	100.0	100.0	–	–	–	–
1k24_A	100.0	100.0	–	–	–	–
1qj8_A	100.0	100.0	–	–	–	–
1i78_A	100.0	100.0	–	–	–	–
2f1v_A	100.0	100.0	–	–	–	–
1p4t_A	100.0	100.0	–	–	–	–
2k0l_A	100.0	100.0	–	–	–	–
2j1n_C	100.0	100.0	–	–	–	–
2qom_A	100.0	100.0	–	–	–	–
1uyo_X	100.0	100.0	–	–	–	–
3jty_A	100.0	100.0	–	–	–	–
2iww_A	100.0	100.0	–	–	–	–
1qd6_D	100.0	100.0	–	–	–	–
1tly_A	100.0	100.0	–	–	–	–

Topology prediction of proteins in the test dataset. BOCTOPUS was run on a randomly selected set of 10 000 among the top performing HMM parameters. Columns 2 and 3 show the fraction correct topology and correct number of predicted strands. For each protein, Columns 4–6 show the location of the most common errors. Strands that are underpredicted (UP) or overpredicted (OP) are shown in Columns 4 and 5, while in Column 6 the strand ids for predicted predicted strands that do not overlap (OV) with the observed location is shown. Column 7 shows the frequency of the most common error among the all mispredictions.

the correct topology for more than one or two of these proteins (Table 4). Given the role played by these atypical TMBs as toxins (Iacovache *et al.*, 2006), further investigation of them in the future will be important.

3.4 The BOCTOPUS web server

The BOCTOPUS web server uses amino acid sequence as input and generates TMB topologies as output. Figure 5 shows an example output from the BOCTOPUS topology prediction pipeline. The lines show the per-residue probabilities for i, M and o regions, respectively. The horizontal bars represent the predicted global topology.

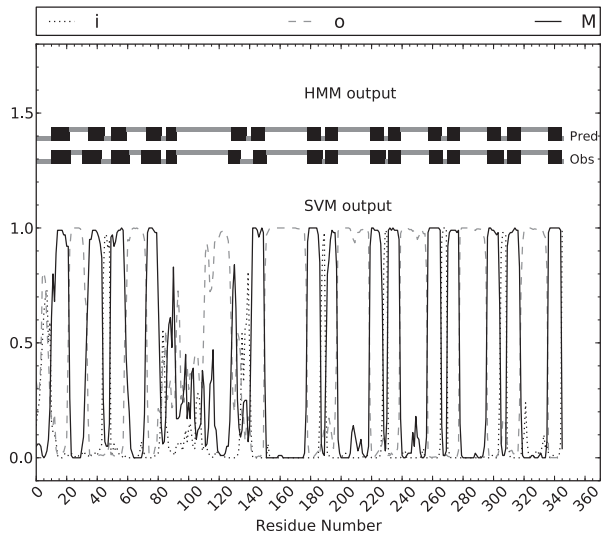

Fig. 5. BOCTOPUS output for Osmoporin OMPC from *E.coli* (2j1nC). The *x*-axis shows the residue number. SVM probability outputs are shown at the bottom (0–1). HMM topology predictions are shown with horizontal bars. Outer-loops, inner-loops and the TM strands are shown in dashed, solid and dotted lines respectively.

Table 4. Multi-chain topology prediction

PDB-ID	Obs. strands	BOCTOPUS	PRED-TMBB	PROFtmb
1tqq_A	4	6	2	6
1wp1_A	4	6	4	–
1yc9_A	4	6	2	4
1uun_A	2	6	4	2
3emo_A	4	4	2	4
7ahl_A	2	10	10	8

Prediction accuracy on multi-chain TMBs. Out of six multi-chain TMBs, BOCTOPUS gets the number of β strands correct only for *Haemophilus influenzae* Hia autotransporter (3emo_A). However, methods in the literature including BOCTOPUS are not optimized to handle multi-chain TMBs.

3.5 TMB identification based on BOCTOPUS predictions

Although BOCTOPUS is best suited for topology prediction and it initially assumes all input sequences to be putative TMBs, we tested its ability to discriminate between TMBs and non-TMBs in a dataset from Freeman and Wimley (2010). Here, a protein is assigned as TMB if the number of predicted strands is larger than a given number N. Figure 6 shows the TMB identification results for different methods combined with BOCTOPUS. The results for the negative set are presented as an average of 10 separate BOCTOPUS runs. For the proteins that are known to be TMBs (positive set), the topology prediction results are taken from the cross-validation test as descried above. Further, the results for PSORTb and other methods mentioned here are obtained from the respective web servers and the inherent homology was not eliminated in prediction. In Table 5, BOCTOPUS(8) and BOCTOPUS(4) refer to cases where the threshold for being a TMB is set at 8 and 4, respectively. BOCTOPUS (8) alone misclassifies 1374 non-TMBs as TMBs, resulting in an MCC value of 0.14. Secondary structure analysis

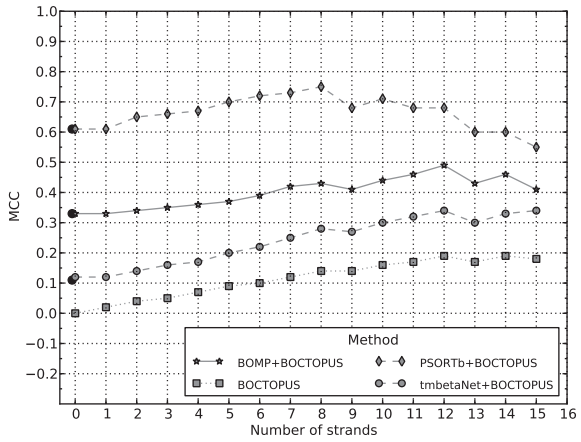

Fig. 6. TMB Identification based on different cutoffs for BOCTOPUS + other methods [performed on a dataset obtained from Freeman and Wimley (2010)]. Black circles represent the MCC value obtained from the method alone.

Table 5. Identification of TMBs

Method	TP	FN	FP	TN	MCC
BOCTOPUS(4)	41	2	5000	8585	0.07
BOCTOPUS(8)	38	5	1374	12 211	0.14
BOMP+BOCTOPUS(4)	36	7	194	13 391	0.36
BOMP+BOCTOPUS(8)	34	9	107	13 478	0.43
BOMP	36	7	238	13 347	0.33
tmbetaNet+BOCTOPUS(4)	39	4	1035	12 550	0.17
tmbetaNet+BOCTOPUS(8)	36	7	342	13 243	0.28
tmbetaNet	41	2	2311	11 274	0.11
PSORTb+BOCTOPUS(4)	36	7	31	13 554	0.67
PSORTb+BOCTOPUS(8)	35	8	16	13 569	0.75
PSORTb	38	5	53	13 532	0.61

TMB identification based on the predicted topology by BOCTOPUS and state of the art methods [performed on dataset obtained from Freeman and Wimley (2010)]. BOCTOPUS (4/8) is the case when sequences are predicted as TMB and the number of predicted strands is ≥ 4 or 8, respectively. Five (out of 48) TMBs in the dataset (Freeman and Wimley, 2010) are toxins and not classified to be transmembrane proteins by OPM (Lomize *et al.*, 2006) and PDBTM (Tusnady *et al.*, 2005) and therefore were excluded from the analysis. Six multi-chain TMBs are included in TMB identification analysis. Highest MCC is obtained by PSORTb + BOCTOPUS(8), where six of the eight FNs are multi-chain TMBs.

of non-TMB proteins predicted as TMB shows that the regions predicted as TM β-strands are enriched in β-strands. Based on DSSP assignment for secondary structure, regions predicted as 'M' have 38.6% residues in β-sheets versus 10.3% for 'I' and 'O' residues.

The methods specialized in identification of TMBs (for example: BOMP, tmbetaNet and PSORTb) are better at identifying TMBs than BOCTOPUS (Table 5). The number of false positives for these methods vary between 53 for PSORTb to 2311 for tmbetaNET. However, when these methods are used together with BOCTOPUS, such that a protein is first predicted by a TMB identification method and then checked by BOCTOPUS if the number of strands is ≥ 4 or 8, the number of false positives is reduced without a large decrease in sensitivity. The highest accuracy is obtained when combining PSORTb with BOCTOPUS(8). The number of false positives is reduced from 53 to 16 with eight losses in TPs. Accordingly, the MCC value increases from 0.61 to 0.75 (Table 5). Six of the eight misclassified TMBs are multi-chain TMBs. This is due to the fact that multi-chain TMBs generally consists of 2–4 β-strands per chain. The other two TMBs misclassified as non-TMBs are Toluene transporter TbuX (3bry_A) and Lipid A deacylase PagL (2erv_A).

4 CONCLUSION

Here, we present an improved topology predictor for TMBs named BOCTOPUS that combines local per-residue predictions with global preferences. BOCTOPUS is based on ideas previously implemented for the topology prediction of HMPs where different residue preference scores derived from sequence profiles are combined to predict the global topology (Jones, 2007; Viklund and Elofsson, 2008). BOCTOPUS is benchmarked on a non-redundant dataset with 36 TMBs with known 3D structure. Based on a 10-fold cross-validation test, the prediction accuracy of BOCTOPUS is higher than earlier methods both when measured on a per-residue and a per-protein basis. BOCTOPUS predicts the correct number of strands in 30 of 36 (83%) TMBs and obtains the correct topology for 70% TMBs in the dataset. We also show that when BOCTOPUS is combined with dedicated TMB identification methods such as BOMP (Berven *et al.*, 2004), PSORTb (Yu *et al.*, 2010) and tmbetaNet (Gromiha *et al.*, 2005b), it can reduce the false positive detection of TMBs. However, the performance in multi-chain TMBs is far from perfect indicating that the correct prediction and identification of such proteins is not a solved problem.

ACKNOWLEDGEMENTS

We thank Drs Paul Horton and Kenichiro Imai for the TMB dataset and for fruitful discussions and Dr Nanjiang Shu for a careful reading of the manuscript.

Funding: Swedish Research Council (VR-NT 2009-5072, VR-M 2010-3555); SSF (the Foundation for Strategic Research); Vinnova through the Vinnova-JSP program; the EU 7th 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.

Berven,F. *et al.* (2004) BOMP: a program to predict integral β-barrel outer membrane proteins encoded within genomes of Gram-negative bacteria. *Nucleic Acids Res.*, **32** (Suppl. 2), W394.

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

Deng,Y. *et al.* (2004) Scoring hidden Markov models to discriminate [beta]-barrel membrane proteins. *Comput. Biol. Chem.*, **28**, 189–194.

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.

Garrow,A. *et al.* (2005) TMB-Hunt: a web server to screen sequence sets for transmembrane β-barrel proteins. *Nucleic Acids Res.*, **33** (Suppl. 2), W188.

Gromiha,M. and Suwa,M. (2006) Discrimination of outer membrane proteins using machine learning algorithms. *Proteins*, **63**, 1031–1037.

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.* (2005a) Application of residue distribution along the sequence for discriminating outer membrane proteins. *Comput. Biol. Chem.*, **29**, 135–142.

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

Hu,J. and Yan,C. (2008) A method for discovering transmembrane beta-barrel proteins in Gram-negative bacterial proteomes. *Comput. Biol. Chem.*, **32**, 298–301.

Iacovache,I. *et al.* (2006) A rivet model for channel formation by aerolysin-like pore-forming toxins. *EMBO J.*, **25**, 457–466.

Jones,D. (2007) Improving the accuracy of transmembrane protein topology prediction using evolutionary information. *Bioinformatics*, **23**, 538.

Kabsch,W. and Sander,C. (1983) *Biopolymers*, **22**, 2577–2637

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.

Matthews,B. (1975) Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451.

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

Ou,Y. *et al.* (2010) Prediction of membrane spanning segments and topology in β-barrel membrane proteins at better accuracy. *J. Comput. Chem.*, **31**, 217–223.

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.

Park,K. *et al.* (2005) Discrimination of outer membrane proteins using support vector machines. *Bioinformatics*, **21**, 4223.

Randall,A. *et al.* (2008) TMBpro: secondary structure,{beta}-contact and tertiary structure prediction of transmembrane {beta}-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.

Remmert,M. *et al.* (2010) Evolution of outer membrane β-barrels from an ancestral $\beta\beta$ hairpin. *Mol. Biol. Evol.*, **27**, 1348.

Rost,B. *et al.* (1994) Redefining the goals of protein secondary structure prediction. *J. Mol. Biol.*, **235**, 13–26.

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

Seshadri,K. *et al.* (1998) Architecture of beta-barrel membrane proteins: analysis of trimeric porins. *Protein Sci.*, **7**, 2026–2032.

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 BBA Proteins Proteomics*, **1814**, 664–670.

Tusnady,G. *et al.* (2005) PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank. *Nucleic Acids Res.*, **33**, D275.

Viklund,H. and Elofsson,A. (2004) Best alpha-helical transmembrane protein topology predictions are achieved using hidden markov models and evolutionary information. *Protein Sci.*, **13**, 1908–1917.

Viklund,H. and Elofsson,A. (2008) OCTOPUS: improving topology prediction by two-track ANN-based preference scores and an extended topological grammar. *Bioinformatics*, **24**, 1662.

Waldispuhl,J. *et al.* (2006) transFold: a web server for predicting the structure and residue contacts of transmembrane beta-barrels. *Nucleic Acids Res.*, **34**, W189.

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

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.

Yu,N. *et al.* (2010) Psortb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics*, **26**, 1608.