Bioinformatics, 32(10), 2016, 1571–1573
doi: 10.1093/bioinformatics/btw025
Advance Access Publication Date: 21 January 2016
Applications Note

Structural bioinformatics

Inclusion of dyad-repeat pattern improves topology prediction of transmembrane β-barrel proteins

Sikander Hayat[1], Christoph Peters[2], Nanjiang Shu[2,3], Konstantinos D. Tsirigos[2] and Arne Elofsson[2]*

[1]Memorial Sloan Kettering Cancer Center, New York City, NY, USA, [2]Stockholm Bioinformatics Center, SciLifeLab, Swedish E-Science Research Center, Stockholm University, Stockholm, SE, 10691, Sweden and [3]Sweden Bioinformatics Infrastructure for Life Sciences (BILS), Stockholm University, Sweden

*To whom correspondence should be addressed.

Associate Editor: Burkhard Rost

Received on 1 July 2015; revised on 14 November 2015; accepted on 14 January 2016

Abstract

Summary: Accurate topology prediction of transmembrane β-barrels is still an open question. Here, we present BOCTOPUS2, an improved topology prediction method for transmembrane β-barrels that can also identify the barrel domain, predict the topology and identify the orientation of residues in transmembrane β-strands. The major novelty of BOCTOPUS2 is the use of the dyad-repeat pattern of lipid and pore facing residues observed in transmembrane β-barrels. In a cross-validation test on a benchmark set of 42 proteins, BOCTOPUS2 predicts the correct topology in 69% of the proteins, an improvement of more than 10% over the best earlier method (BOCTOPUS) and in addition, it produces significantly fewer erroneous predictions on non-transmembrane β-barrel proteins.

Availability and implementation: BOCTOPUS2 webserver along with full dataset and source code is available at http://boctopus.bioinfo.se/

Contact: arne@bioinfo.se

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

1 Introduction

Transmembrane β-barrels (TMBs) are located in the outer membranes of bacteria, chloroplasts and mitochondria and are involved in transport of substrates across the cell membrane. Also, they are key components of translocation complexes. Bacterial TMBs consist of an even number of anti-parallel β-strands that are surrounded by long outer-loops and short inner-loops. Eukaryotic TMBs can have an odd number of strands. The residues located in transmembrane β-strands follow a dyad-repeat pattern, where alternating residues face the lipids and the pore region (Wimley, 2003), respectively. BOCTOPUS (Hayat and Elofsson, 2012), a previously published two-stage TMB topology predictor, did not predict the lipid/pore-facing orientation of residues in the transmembrane β-strands but BOCTOPUS2 does. The use of this information improves topology predictions.

2 Methods

Forty-two TMB sequences, culled at 20% sequence identity using the PISCES server (Wang and Dunbrack, 2005) are divided into 10 subsets for cross-validation (Supplementary Table S1). To ensure that no homology information is used, all proteins from the same super-family are put into the same cross-validation group. Membrane boundaries, 3D structures and super-family classification are obtained from the OPM database (Lomize *et al.*, 2006). Residues in the transmembrane strands are labelled as pore or lipid-facing based on their side-chain orientation relative to the barrel center. In addition, a dataset (also reduced to 20% sequence identity) consisting of 370 (28 eukaryotic and 342 prokaryotic) known TMBs and 7558 non-TMBs that contain a signal peptide was downloaded from Uniprot (Consortium *et al.*, 2015).

BOCTOPUS2 consists of three layers (Fig. 1): First, the preferences and the class-probability for each residue to be in each state (i.e. outer-loop (*o*), inner-loop (*i*), pore-facing (*p*) or lipid-facing (*l*) β-strand region) are predicted using a Support Vector Machine (SVM). Secondly a filter is applied to detect the barrel region. These class-probabilities are used to create a ($L \times 4$) protein profile, where L is the number of residues in the input protein sequence. We then apply a filter to estimate the barrel region. In this step, we use the sum of values for the in-facing and out-facing classes over a window of 100 residues centered at the residue in question to predict the barrel region. The region where this sum is $\geq$0.3 is considered the barrel region. Residues outside this region are considered to be in the non-barrel region and all lipid-facing (*l*) and pore-facing (*p*) probability values in that region are set to 0 and all inner-loop (*i*) values are set to 1, so as to suppress a false transmembrane beta-strand prediction. In the final step, the updated protein profile obtained after applying the filter is used as the input to a Hidden Markov-like Model to predict the final global topology.

The advantage of this architecture is twofold; the SVM can use a window of residues to estimate the localization of a single residue, and the Hidden Markov Model does not require training as all transition probabilities can be set to 1. This is in contrast to BOCTOPUS where parameters had to be optimized. For each protein, a multiple sequence alignment is generated using 4 iterations of HHblits against Uniprot20 database (Remmert *et al.*, 2011). Sequences with more than 60% gaps are filtered out. A position specific substitution matrix (PSSM) is then generated using blastpgp (Altschul *et al.*, 1997) and used as input to the next step. Optimal results were obtained using window sizes of 9, 13, 23 and 31 residues for predicting lipid-facing, pore-facing, inner and outer loops, respectively.

2.1 Filter for detecting non-TMB regions

Some TMBs contain long non-barrel regions and sporadic strand predictions can occur in these proteins causing errors in topology predictions. Earlier methods have often been trained on PDB sequences that exclude these regions making these methods less suitable for large-scale predictions. Here, we use the observation that β-strands incorrectly predicted in the non-barrel region are distantly separated from each other. A running average of the *l* and *p* probabilities (Fig. 1) over a window size of 100 is summed and an empirical threshold value of $\geq$0.3 is used to identify the transmembrane barrel region.

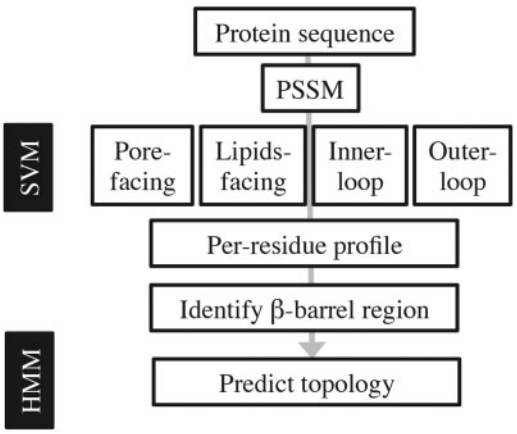

Fig. 1. BOCTOPUS2 pipeline – HHblits is used to obtain a position specific scoring matrix (PSSM), 4 SVMs are used to predict the location of each residue. A per-residue profile is constructed by combining the output probabilities of SVMs. Then a filter is applied to detect the likely transmembrane β-barrel region. The resulting per-residue profile is used as an input to a Hidden Markov Model to predict the topology

2.2 HMM-architecture

The HMM consists of states for inner- and outer-loops, and lipid- and pore-facing membrane strand residues. The architecture initially used in BOCTOPUS is simplified to reduce the number of parameters (Supplementary Fig. S1). The number of residues in a β-strand can be either 9 or 11. HMM parameters are independent of the dataset as all transition and emission probabilities in a given state are set to 1. The Viterbi algorithm is subsequently used to find the optimal topology.

3 Results

BOCTOPUS2 predicts the correct topology for 29 proteins out of 42 (69%) proteins (Table 1). This is better than other methods and, in addition, the performances of all earlier methods are possibly overestimated as some of the 42 proteins have been used in their training sets (Supplementary Information). One main reason for the improved results is that the filter in BOCTOPUS2 avoids incorrect prediction of strands in the non-barrel region. This filter also aids discrimination between TMBs and non-TMBs. In a dataset of 7558 non-TMBs and 370 TMBs BOCTOPUS2 (Freeman and Wimley, 2010) predicts three or more strands in 81% of the TMBs at a specificity of 98.9% (Supplementary Table S2). Amongst the methods tested here, only HHomp (Remmert *et al.*, 2009) and BETAWARE have a higher Mathews correlation coefficient (MCC) for barrel/non-barrel discrimination (Supplementary Table S2). The detection of eukaryotic TMBs seems to be more difficult as BOCTOPUS2 only detects 10 out of 28, possibly due to different evolutionary origins (Supplementary Table S2).

Table 1. Topology prediction and discrimination accuracy comparison

	Topology prediction					Discrimination	
Method	Topology	TM	Q2	Q3	SOV	TPR	FPR
BOCTOPUS2	29	35	91	89	94	81.4	1.1
BOCTOPUS	24	29	92	85	87	88.7	11.5
profTMB	20	20	86	–	64	82.9	11.2
PRED-TMBB	11	16	83	76	71	66.1	10.9
BETAWARE	7	18	82	74	75	89.9	0.7
HHomp	–	–	–	–	–	84.7	0.2
BOMP	–	–	–	–	–	71.9	5.2
F–W β-barrel Analyzer	–	–	–	–	–	88.2	16.6

Comparison of topology prediction methods. Topology - Number of proteins with the correct topology. TM – Number of proteins with correct number of predicted strands. SOV is the segment overlap (Rost *et al.*, 1994), Q2 is the two-state (membrane/not-membrane) prediction accuracy and Q3 is the three-state (membrane, inner-loop, outer-loop) prediction accuracy. Q4, the four state (lipid-facing, pore-facing, inner-loop, outer-loop) accuracy for BOCTOPUS2 is 89. BOCTOPUS2 results are cross-validated while all other methods have included some of the 42 proteins in their training set. The two last columns TPR and FPR describe the true and false positive rate in the discrimination dataset. BOCTOPUS (Hayat and Elofsson, 2012), profTMB (Bigelow and Rost, 2006), PRED-TMBB (Bagos *et al.*, 2004), BETAWARE (Savojardo *et al.*, 2013), HHomp (Remmert *et al.*, 2009) and F-W β-barrel Analyzer (Freeman and Wimley, 2010) results were obtained using standalone versions.

4 Conclusions

We present BOCTOPUS2, an improved topology predictor for TMBs that exploits the dyad-repeat feature of bacterial TMBs to identify transmembrane β-strands. BOCTOPUS2 can also identify the transmembrane barrel domain in long sequences and predict the correct topology in ~69% proteins in our dataset. Further, BOCTOPUS2 predicts the lipid-facing or pore-facing status of residues in identified β-strands. We have recently shown that accurate topology and residue-orientation prediction is beneficial for 3D modelling of TMBs (Hayat *et al.*, 2015).

Funding

This work was supported by grants from the Swedish Research Council (VR-NT 2012-5046), SSF and Vinnova through the Vinnova-JSP program.

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 {beta}-barrel outer membrane proteins. *Nucleic Acids Res.*, **32**, W400.

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

Consortium, U. *et al.* (2015) Uniprot: a hub for protein information. *Nucleic Acids Res.*, **43**, D204.

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

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

Hayat,S. *et al.* (2015) All-atom 3d structure prediction of transmembrane β-barrel proteins from sequences. *Proc. Natl Acad. Sci. USA*, **112**, 5413–5418.

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

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

Remmert, M. *et al.* (2011) Hhblits: lightning-fast iterative protein sequence searching by hmm-hmm alignment. *Nat. Methods*, **9**, 173–175.

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

Savojardo,C. *et al.* (2013) Betaware: a machine-learning tool to detect and predict transmembrane beta barrel proteins in prokaryotes. *Bioinformatics*, bts728.

Wang,G. and Dunbrack,R.L. (2005) Pisces: recent improvements to a pdb sequence culling server. *Nucleic Acids Res.*, **33**, W94–W98.

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