W186–W188 *Nucleic Acids Research, 2006, Vol. 34, Web Server issue*
doi:10.1093/nar/gkl262

PROFtmb: a web server for predicting bacterial transmembrane beta barrel proteins

Henry Bigelow[1,2,*] and Burkhard Rost[1,2,3]

[1]CUBIC, Department of Biochemistry and Molecular Biophysics, Columbia University, 630 West 168th Street, New York, NY 10032, USA, [2]Columbia University Center for Computational Biology and Bioinformatics (C2B2), 1130 St. Nicholas Avenue Rm. 802, New York, NY 10032, USA and [3]North East Structural Genomics Center (NESG), Irvine Cancer Center, 1130 St. Nicholas Avenue Rm. 802, New York, NY 10032, USA

Received February 14, 2006; Revised March 1, 2006; Accepted March 31, 2006

ABSTRACT

PROFtmb predicts transmembrane beta-barrel (TMB) proteins in Gram-negative bacteria. For each query protein, PROFtmb provides both a *Z*-value indicating that the protein actually contains a membrane barrel, and a four-state per-residue labeling of upward- and downward-facing strands, periplasmic hairpins and extracellular loops. While most users submit individual proteins known to contain TMBs, some groups submit entire proteomes to screen for potential TMBs. Response time is about 4 min for a 500-residue protein. PROFtmb is a profile-based Hidden Markov Model (HMM) with an architecture mirroring the structure of TMBs. The per-residue accuracy on the 8-fold cross-validated testing set is 86% while whole-protein discrimination accuracy was 70 at 60% coverage. The PROFtmb web server includes all source code, training data and whole-proteome predictions from 78 Gram-negative bacterial genomes and is available freely and without registration at http://rostlab.org/services/proftmb.

INTRODUCTION

Transmembrane beta-barrel (TMB) proteins form a beta-barrel as a single beta-sheet joined at its edges. The sheet is 'all-next-neighbor'(1), meaning all paired strands are adjacent in sequence. N- and C-termini of TMBs always reside in the periplasm. The architecture can be described as the repeating pattern, where 'up' means towards the extracellular side: N-term, [up-strand, outer loop, down-strand, periplasmic hairpin]$_n$, C-term. PROFtmb, originally published in (2) provides a prediction of residues in these four states (example Figure 1). It exploits statistical features of TMBs including enrichment of beta- and gamma-hairpins in the periplasm, lengths of outer loops, 'aromatic cuffs' and the 'hydrophobic belt', and follows several design ideas from other Hidden Markov Model (HMM)-based TMB predictors (3,4). PROFtmb predicts TMBs from Gram-negative bacteria only. It does not predict TMBs from mitochondria, chloroplasts or the outer membranes of 'atypical' Gram-positive bacteria called mycolata, which have thicker mycolic acid containing outer membranes.

PROCEDURE AND EXAMPLE OUTPUT

Users submit one or more FASTA-formatted protein sequences. For each sequence, PROFtmb builds a PSI-BLAST profile and runs the prediction, attempting to find the best fit of the protein to its TMB-based architecture, indicated as a *Z*-value. Results are always returned on a webpage, and take $\sim$4 min per 500-residue protein. In the case of more than one sequence, an email of the results URL is sent.

If the query protein receives *Z*-value $\geq$ 4.0, PROFtmb provides a four-state (upward-strand, downward-strand, outer loop, periplasmic loop) per-residue prediction. Graphical output consists of color-coded four state posterior probability plots and amino acid sequence (Figure 1). Amino acid color indicates the final prediction, and usually corresponds to the state with maximum posterior probability, but with 'corrections' based on context shown with lighter-weight font [described in 'Decoding' section of the Supplementary Data of (2)]. While we did not quantify confidence levels for per-residue prediction, higher *Z*-values tend to have fewer corrected residues and greater contrast in state posterior probabilities.

In the example shown (Figure 1), OMPA from *Escherichia coli* [PDB: 1g90 (5) chain A] is predicted correctly at high confidence as an eight-stranded TMB. This result is expected, given PROFtmb was trained on very similar sequences. In most predictions on real TMBs, corrected residues are only

*To whom correspondence should be addressed. Tel: +1 212 851 4669; Fax: +1 212 851 5176; Email: hrbigelow@gmail.com

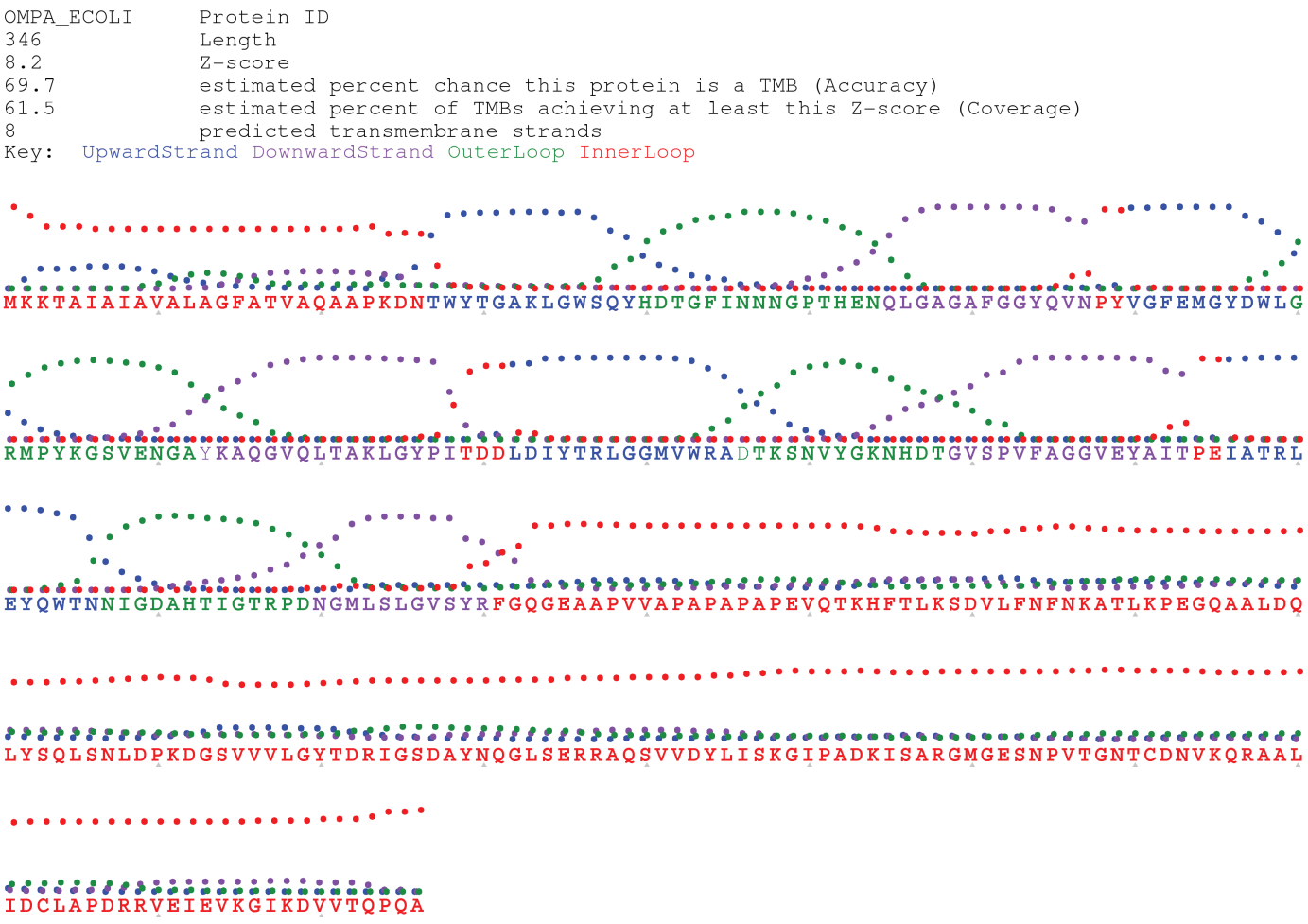

Figure 1. True positive output example. PROFtmb prediction for OMPA from *E.coli* [PDB: 1g90 (5) chain A], a true TMB. Note that predicted strands have high contrast between state probabilities for a majority of their length.

found at the boundaries between strands and loops. Also, most strand and loop states have the best state close to probability 1.

In the second example shown (Figure 2), heme acquisition system protein A from *Serratia marcescens*, of the gammaproteobacteria class (Gram-negative) illustrates a false positive prediction. It receives a low but above-threshold *Z*-value of 4.8. In fact, the structure [PDB: 1B2V (6)] consists of a seven-stranded beta-sheet against four α-helices. PROFtmb does correctly predict the locations of five of the strands. Notice that predicted strands four, five and six have poor contrast in posterior probability, indicating a poor fit to the PROFtmb model.

Finally, proteins shorter than 140 or longer than 1392 residues receive *Z*-value −10 000 (data not shown). The lower length of 140 is a conservative estimate of the smallest possible TMB, while the upper bound reflects the limit of our test set for *Z*-value calibration.

Occasionally, PROFtmb will assign *Z*-value less than four to a known TMB. Unfortunately, in such a case, the fact that it is a TMB can't be used to help produce a reliable per-residue prediction since PROFtmb derives the prediction from sequence alone. This occurred in about 15% of the cases in our test set (see 'Performance Evaluation' in 'Methods' tab on website).

DISCUSSION

In our original paper (2) we used PSI-BLAST profiles run with options −h 1 (*E*-value cutoff for inclusion in next pass) and −j 2 (number of iterations), and did not explore the effect of different profiles on PROFtmb accuracy, either for whole protein or per-residue prediction. Since then, we have run 8-fold jackknife tests (leave one out, seven in) on the original SWISS-PROT sequence versions of eight PDB structures (SetTMBfull: 1a0s_P, 1af6_A, 1bt9_A, 1fep_A, 1prn, 1qd5_A, 1qj9_A, 1qjp_A). We built sets of PSI-BLAST profiles with 30 different combinations of settings −h {1, 0.1, 0.01, 0.001, 0.0001, 0.00001} and −j {2, 3, 4, 5, 6} and used each set in a separate jackknifed test. The original Q2 accuracy, with settings −h 1 −j 2 was 86.0%, while the best settings, −h 0.0001 −j 2 achieved 87.3% Q2 accuracy. As a result, we changed the defaults to −h 0.0001 −j 2. Additionally, we now allow the user to select these parameters. We have not estimated the effects of PSI-BLAST settings on whole-protein prediction yet. Currently, *Z*-value and resulting estimated accuracy and coverage are calibrated from our

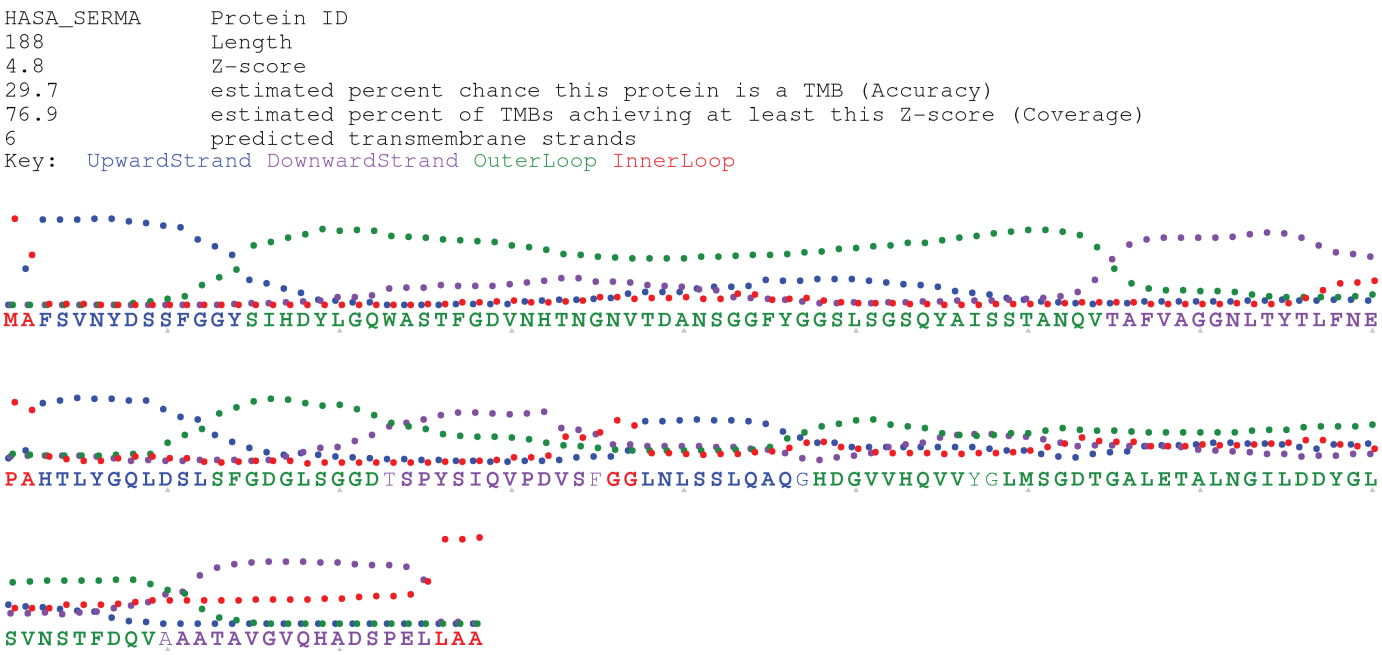

Figure 2. False positive output example. Heme acquisition system protein A (HasA) from *Serratia marcescens* [PDB: 1B2V (6)], a secreted hemophore with architecture beta-alpha-beta (6)-alpha (2) according to SCOP (7). Predicted strands four, five and six show poor contrast in state probabilities and indicate a poor fit to the model.

original sequence-unique set called SetROC, containing a representative set of proteins from SWISS-PROT. As sequence databases are updated, we will periodically re-calibrate *Z*-values. A cluster plot and resulting accuracy versus coverage curve can be found in the 'Methods' section of the website.

DOWNLOADS

Predictions on 78 Gram-negative proteomes are available in the Download section, updated since original publication as follows. First, length-adjusted bits score was replaced by *Z*-value, which gives slightly improved discrimination on our test set (unpublished data). Second, per-residue predictions were re-run using updated PSI-BLAST profiles, with option −h 0.0001 rather than −h 1. Both changes are expected improvements, but haven't been rigorously tested. Third, the model architecture now explicitly includes BEGIN and END states, representing the beginning and end of the amino acid sequence. This is required for the current version of the software.

The PROFtmb software is a general profile-HMM allowing specification of model architecture, encoding and decoding. The training data, consisting of eight TMB sequences with hand-annotated per-residue labeling based on their 3D structures, is available as well. Interested users may download and compile the C++ source code and use PROFtmb with the original training data or modify it. We make it available in the spirit of reproducibility, and encourage interested readers to contact the authors for more detailed advice.

ACKNOWLEDGEMENTS

Thanks to Pier Luigi Martelli and Pantelis Bagos for helpful discussions, generous use of data and sharing unpublished ideas. Thanks to Amos Bairoch (SIB, Geneva), Rolf Apweiler (EBI, Hinxton), Phil Bourne (San Diego University), and their crews for maintaining excellent databases and to all experimentalists who enabled this analysis by making their data publicly available. Last, not least, thanks to all those who deposit their experimental data in public databases, and to others who maintain these databases. This work was supported by grant R01-LM07329-01 from the National Library of Medicine. Funding to pay the Open Access publication charges for this article was provided by NIH/NLM R01-LM07329-01.

Conflict of interest statement. None declared.

REFERENCES

1. Schulz,G.E. (2003) Transmembrane beta-barrel proteins. *Adv. Protein Chem.*, **63**, 47–70.
2. Bigelow,H.R., Petrey,D.S., Liu,J., Przybylski,D. and Rost,B. (2004) Predicting transmembrane beta-barrels in proteomes. *Nucleic Acids Res.*, **32**, 2566–2577.
3. Bagos,P.G., Liakopoulos,T.D., Spyropoulos,I.C. and Hamodrakas,S.J. (2004) A Hidden Markov Model method, capable of predicting and discriminating beta-barrel outer membrane proteins. *BMC Bioinformatics*, **5**, 29.
4. Martelli,P.L., 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.
5. Arora,A., Abildgaard,F., Bushweller,J.H. and Tamm,L.K. (2001) Structure of outer membrane protein A transmembrane domain by NMR spectroscopy. *Nature Struct. Biol.*, **8**, 334–338.
6. Arnoux,P., Haser,R., Izadi,N., Lecroisey,A., Delepierre,M., Wandersman,C. and Czjzek,M. (1999) The crystal structure of HasA, a hemophore secreted by *Serratia marcescens*. *Nature Struct. Biol.*, **6**, 516–520.
7. Lo Conte,L., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2002) SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res.*, **30**, 264–267.