Beta barrel trans-membrane proteins: Enhanced prediction using a Bayesian approach

Paul D. Taylor[1], Christopher P. Toseland[2], Teresa K. Attwood[3] and Darren R. Flower[1*]
[1]The Jenner Institute, University of Oxford, Compton, Newbury, Berkshire, RG20 7NN, UK; [2]National Institute for Medical Research, The Ridgeway, Mill Hill, London, NW7 1AA, UK; [3]Faculty of Life Sciences & School of Computer Science, The University of Manchester, Oxford Road, Manchester M13 9PT, UK;
Darren R. Flower* - Email: darren.flower@jenner.ac.uk; Phone: +44 1635 577954; Fax: +44 1635 577908;
* Corresponding author

received October 04, 2006; accepted October 06, 2006; published online October 07, 2006

Abstract:

Membrane proteins, which constitute approximately 20% of most genomes, form two main classes: alpha helical and beta barrel transmembrane proteins. Using methods based on Bayesian Networks, a powerful approach for statistical inference, we have sought to address β-barrel topology prediction. The β-barrel topology predictor reports individual strand accuracies of 88.6%. The method outlined here represents a potentially important advance in the computational determination of membrane protein topology.

Keywords: beta barrel transmembrane protein; prokaryotic membrane proteins; Bayesian Networks; prediction method; sub-cellular location

Background:

Accurate and reliable prediction of protein structure and function remains a challenge. Of particular importance is the prediction of membrane proteins, as, unlike soluble and fibrous proteins, membrane proteins remain poorly tractable targets for the principal experimental methods of structure determination: X-ray crystallography and multidimensional nuclear magnetic resonance (NMR) spectroscopy. **[1]** This problem is highlighted by the observation that 20% of most genomes encode membrane proteins **[2]**, yet the number of solved membrane protein structures is approximately 2% of the RCSB Protein Data Bank (PDB). **[3, 4]**

Membrane proteins fall into two structural classes: α-helical and β-barrel. At present, the only known location for TM β-barrels is the outer membrane of Gram-negative bacteria. **[5]** Although there is very strong evidence for their presence in mitochondrial and chloroplast membranes (e.g., the voltage-dependent anion channel (VDAC), the long-standing homologue candidate in the outer mitochondrial membrane). The SCOP database classifies TM β-barrels into 6 structural superfamiles: OmpA-like, OmpT-like, OmpLA, porins, TolC and Leukocidin (α Haemolysin) [6]. β-barrels have been shown to have a variety of functions, including the passive transport of ions and small hydrophilic molecules, the export of xenobiotics, import of siderophore-bound iron, and a role in bacterial pathogenicity. **[7-10]** Despite these widely different functions, these proteins show a remarkable degree of structural similarity, which has led Schulz to identify 8 rules summarising β-barrel construction. **[5]** Of these, two are of particular importance when attempting to predict TM β-barrel topology: rule two states that both the N- and C-termini are at the periplasmic end of the barrel, restricting the strand number to even values; and rule 4, that external β-strand connections are long loops (termed L1, L2, etc.), whereas the periplasmic strand connections are generally short (T1, T2, etc.).

Although the 8 rules defined by Schulz characterise β -barrel construction well, the prediction of barrel topology from sequence remains a difficult task owing to several complicating factors. First, identifying potential TM strands as stretches of sequence where residues alternate between polar and non-polar grossly over-simplifies the problem, as this pattern is frequently broken by non-polar residues on the interior of the barrel. Second, the average length of β-strands is seldom more than half a dozen residues; they are therefore much harder to distinguish than longer TM α-helices. **[11]** Finally, the most significant hindrance to β-barrel topology prediction is probably the lack of solved structures on which to train predictive methods. **[12]**

This paper describes the construction of a predictor for a beta-barrel membrane protein topology, based on machine learning Bayesian Networks (BNs). BNs are considered especially suited to computational biology, as they provide a flexible and powerful framework for statistical inference, and learn model parameters from data. **[13]**

Methodology:

Data-set

A dataset of TM proteins of experimentally verified α-helical topology were required to train the method. The data-set was obtained from the TMPDB database (release 6.2) **[12]**, in which topologies have been determined using X-ray crystallography, NMR, gene fusion, substituted cysteine accessibility, *N*-linked glycosylation experiment and other biochemical methods. Non-redundant subsets of TMPDB were used, and hence sequence similarity between proteins was less than 30%. The β-barrel data-set (TMPDB_β_non-redundant) consisted of 15 proteins.

Beta-Barrel Transmembrane Protein Topology Predictor

A static full Bayesian model was used as it best fulfils the requirements expected of the network. The main advantage of such a model, compared with its naïve counterpart, is that the output probability is not a product of probabilities from each descriptor, but a model which associates one probability with each combination of descriptors. Thus, overall performance is typically better, but never worse, than that of the best individual descriptor.

Descriptors of amino acid properties were used to characterise amino acids. Each descriptor was averaged, using a sliding-window, to produce a set of meta-descriptors. Each meta-descriptor represents one node, the state of which influences, probabilistically, the state of the output node. There are two main facets to the descriptor-based sliding-window methodology: the use of appropriate descriptors that will provide inferences about whether a residue is TM or not; and the use of a sliding window as a mechanism for taking into consideration the surrounding environment of a residue.

The descriptors used were the 434 amino acid property scales in the AAIndex database (release 6.0). **[14]** The scales provide a large range of amino acid properties, including: size, charge, hydrophobicity and more recondite propensities (such as membrane buried preference parameter). The initial step in the creation of the methods was the production of a temporary meta-training-set. For every residue in the training-set, the 434 sliding-window-averaged values (one for each scale) are calculated, and whether or not the residue is TM located is recorded. The BN is trained on this meta-data-set. The optimal shape of the sliding window was found to be trapezoid (data not shown), with 50% less weighting of the two residues at either end – this is because these residues are further from the residue being analysed, and hence have less influence. However, this may not account for possible tertiary conformations that may bring residues close together that are far apart in the sequence. For TM β-strands, a window size of 7 residues was used. The network was then trained on the sliding-window values and residue location. During training, the network attempts to find which descriptor values best correlate with the residue occurring in a TM region.

When presented with a test sequence, prediction initially follows the same process as training, with each residue assigned a sliding-window-averaged value for each descriptor. The network then moves through the sequence, and determines whether the values encountered are typical of a TM-located residue. As prediction is made on an individual-residue basis, there is a requirement for post-network processing to translate the prediction from single residues to TM regions. This is done with reference to knowledge of TM-region tendencies observed in solved and/or well characterised structures. Accordingly, post-network processing imposes the following rules upon possible TM regions: β-strands must have a minimum length of 6 and a maximum length of 25 residues. Short strand predictions are disregarded and long predictions are split at the central residue and the two residues either side of the split are designated non-transmembrane.

Results and Discussion:

A BN, predictive of β-strands in TM proteins, was constructed and its accuracy assessed using 15 proteins taken from a non-redundant data-set of experimentally verified topology. The β-strand predictor initially appeared to produce disappointing results when considering overall protein topology accuracy (42.7%). When considering individual β-strands, however, the accuracy of prediction was found to be 88.6%: much higher than the relatively low topology accuracy. This clearly shows that the method can accurately distinguish strands from non-membranous regions of the protein; a task that, as discussed earlier, presents many challenges, owing to the short and variable nature of β-strands. In common with problems reported for other predictors **[15]**, the predictions were a little shorter than the actual strand, the predicted length being on average 92.6% of the actual value.

The β-strand predictor showed low overall topology accuracy, but high strand accuracy. The major observed failing of the method was to predict two separate strands as one combined or double strand. These false double strands fell below the maximum 25-residue strand length threshold and were thus not split into two separate strands. The problem arises only between strands separated by short intra-cellular turns. The method assumes the short span of turn residues to be part of a strand, as they are surrounded by longer stretches of strand residues. This failing may, in part, be caused by the sliding-window method, which takes into account the surrounding environment and is therefore less sensitive to anomalous short regions. The topological consequence of such combined strands is barrel predictions with odd strand numbers. As all β-barrels have even numbers of strands, this error is easily spotted yet is surprisingly difficult to correct. Algorithms attempting to identify "double strands" are unable to distinguish them easily from long strands. This problem is exemplified by the Ferrichrome iron receptor (FHUA) protein from *E. coli*, which has the shortest and longest strand lengths in the data-set: 7 and 24 residues respectively. If we predict an odd strand number for this protein, we are required to distinguish a double strand, which may be as short as 17 residues, from a long strand of up to 24 residues. This problem arises, in part, from the very small training-set used to train the network. Unusually long or unusually short strands, as compared with the average strand length, are relatively infrequent; thus, any network would have few training examples on which to base its predictions. Although the accuracy of the β-strand predictor suffers from the use of a small data-set, and would clearly benefit from re-training as more structures become available, its overall predictive power nevertheless compares very favourably with methods developed using other artificial intelligence techniques.

Conclusion:

The method described here represents an important advance in the computational determination of membrane protein structural class and topology. The beta-barrel TM protein topology predictor has good accuracy. The method described offers a useful and complementary tool for the analysis of membrane proteins for a wide range of possible applications.

Acknowledgement:

PDT wishes to thank the MRC for a priority area studentship. The Jenner Institute (Formally, The Edward Jenner Institute for Vaccine Research) wishes to thank its sponsors: GlaxoSmithKline, the Medical Research Council, the Biotechnology and Biological Sciences Research Council, and the UK Department of Health.

References:

[01] A. Arora & L. K. Tamm, *Curr. Opin. Struct. Biol.,* 11:540 (2001) [PMID: 11785753]

[02] J. Liu & B. Rost, *Protein Sci.,* 10:1970 (2001) [PMID: 11567088]

[03] H. M. Berman, *et al., Acta. Crystallogr. D. Biol. Crystallogr.,* 58:899 (2002) [PMID: 12037327]

[04] R. Casadio, *et al., Brief Bioinform.,* 4:341 (2003) [PMID: 14725347]

[05] G. E. Schulz, *Curr. Opin. Struct. Biol.,* 10:443 (2000) [PMID: 10981633]

[06] A. G Murzin, *et al., J. Mol. Biol.,* 247:536 (1995) [PMID: 7723011]

[07] G. E. Schulz, *Curr. Opin. Struct. Biol.,* 6:485 (1996) [PMID: 8794162]

[08] V. Koronakis, *et al., Nature,* 405:914 (2000) [PMID: 10879525]

[09] A. Pautsch & G. E. Schulz, *Nat. Struct. Biol.,* 5:1013 (1998) [PMID: 9808047]

[10] J. Vogt & G. E. Schulz, *Structure Fold Des.,* 7: 1301 (1999) [PMID: 10545325]

[11] Y. Mandel-Gutfreund & L. M. Gregoret, *J. Mol. Biol.,* 323:453 (2002) [PMID: 12381301]

[12] M. Ikeda, *et al., Nucleic Acids Res.,* 31:406 (2003) [PMID: 12520035]

[13] Pearl, J, Probablistic Reasoning in Intellegent Systems: Networks of Plausible Inference. (1988), San Mateo, California: Morgan Kaufman.

[14] J. L. Gardy, *et al., Nucleic Acids Res.,* 31:3613 (2003) [PMID: 12824378]

[15] D. M. Lao, *et al., Bioinformatics,* 18:1562 (2002) [PMID:12490439]

Edited by P. Kangueane
Citation: Taylor *et al.*, Bioinformation 1(6): 231-233 (2006)