Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Prediction Model**

Alpha helical trans-membrane proteins: Enhanced prediction using a Bayesian approach

Paul D. Taylor[1], Christopher P. Toseand[2], Teresa K. Attwood[3] and DarrenR.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 06, 2006; accepted November 01, 2006; published online November 14, 2006

Abstract:

Membrane proteins, which constitute approximately 20% of most genomes, are poorly tractable targets for experimental structure determination, thus analysis by prediction and modelling makes an important contribution to their on-going study. Membrane proteins form two main classes: alpha helical and beta barrel trans-membrane proteins. By using a method based on Bayesian Networks, which provides a flexible and powerful framework for statistical inference, we addressed α-helical topology prediction. This method has accuracies of 77.4% for prokaryotic proteins and 61.4% for eukaryotic proteins. The method described here represents an important advance in the computational determination of membrane protein topology and offers a useful, and complementary, tool for the analysis of membrane proteins for a range of applications.

Keywords: trans-membrane protein; alpha helix; static full Bayesian model; prediction; amino acid descriptors

Background:

Membrane proteins, which are poorly tractable targets for the main experimental methods of structure determination - X-ray crystallography and multidimensional nuclear magnetic resonance (NMR) spectroscopy - yet form approximately 20% of most genomes **[1-4]**, fall into two structural classes: α-helical and β-barrel. α-helical membrane proteins are responsible for interactions between most cells and their environment. **[5]** Trans-membrane (TM) helices are typically encoded by stretches of 17-25 residues **[6]**, which provide sufficient length to cross the membrane. **[7]** A compositional bias towards hydrophobic residues is apparent in the TM helices, as they must make complementary interactions with the hydrophobic lipid bilayer. **[8]** α-helical proteins vary in topology, from single TM regions to "serpentine" structures consisting of over 20 TM helices, which are separated by hydrophilic regions that loop alternately in and out of the extra-cellular space and the cytoplasm. **[9]** The functions that have been observed for α-helical proteins are as varied as their topologies, including signal recognition, receptors, transfer of molecules and ions across the membrane, and energy translocation and conservation. **[10-12]** The function of membrane proteins with multiple α-helices, where TM domains often combine to form a tightly-coupled structure, is dependent on their final 3-dimensional conformation. **[13]** Consequently, the number and arrangement of TM domains is often conserved within a protein family.

We describe here construction of a predictor of α-helical membrane proteins. This method is based on 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. **[14]**

Methodology:

Data-sets

A dataset of TM proteins of confirmed α-helical topology was required for training the method. This data-set was taken from the TMPDB database (release 6.2) **[15]**, in which topologies have been determined using a variety of experimental techniques such as X-ray crystallography. Non-redundant subsets of TMPDB were created where sequence similarity between proteins was less than 30%. The data-set (TMPDB_α_non-redundant) comprised 231 proteins, of which 93 were from eukaryotic and 138 from prokaryotic organisms.

Predictor

A static full Bayesian model was used for the network. The principal advantage of a static full Bayesian model, compared with a naïve model, is that the output probability is not a product of probabilities from each descriptor. Rather, a full Bayesian model associates one probability with each combination of descriptors. Thus, overall performance is at least as good as that of the best individual descriptor.

434 amino acid property scales were used as descriptors during model building. The descriptors were obtained from the AAIndex database. **[16]** The scales provide a large range of amino acid properties, including: size, charge, hydrophobicity and propensities (such as membrane buried preference parameter and surface accessibility). Each descriptor was averaged, using a sliding-window, to produce a set of meta-descriptors. The environment surrounding the residue is taken into

ISSN 0973-2063
Bioinformation 1(6): 234-236 (2006)

account through this method. The structure of the network requires that each meta-descriptor represents one node, the state of which influences, probabilistically, the state of the output node. Thus our methodology seeks to identify appropriate descriptors correlated with trans-membrane burial and to exploit a sliding window which implicitly considers neighbourhood effects from the surrounding environment.

The initial step in our method development cycle 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) were calculated, and whether or not the residue was part of a TM region was recorded. The BN was then trained with this meta-training-set. A window size of 13 residues was used. The optimal shape of the sliding window for these particular predictors was found to be trapezoid (data not shown), with 50% less weighting of the two residues at either end. This may result from the reduced influence of distal residues relative to the immediate environment. Naturally, this does not account for possible tertiary conformations that may bring residues close together that are far apart in the sequence. During training, the network attempts to find which descriptor values best correlate with residues in a TM region.

Prediction by the network, when presented with a test sequence, initially follows the same process as training, each residue in the sequence being assigned sliding-window-averaged values for all the descriptors. The network then moves through the sequence, and determines whether the values for the descriptors 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 well characterised structures. Accordingly, post-network processing imposes the following rules: α-helices cannot be shorter than 14 residues or longer than 40 residues. Thus short helix predictions are disregarded and those >40 residues are split into two at the most hydrophilic of the central 5 residues. Three residues either side of the split point are labelled as non-trans-membrane.

Results and Discussion:

A BN, predictive of α-helices in TM proteins, was constructed and its accuracy assessed using leave-one-out and leave-five-out cross validation with a non-redundant protein data-set consisting of 231 prokaryotic and eukaryotic proteins of known structure. The results are summarised in Table 1. Correct prediction of topology is defined as correct prediction of the number of TM domains, the correct prediction of location of the central residue to an accuracy of 5 residues either side, and the correct prediction of N-terminal location.

Although the results show a good level of topology prediction accuracy, comparable with the best publicly available method, for both eukaryotic and prokaryotic sequences, there is an obvious difference in performance for the two. In order to address this, the average number and length of TM segments for the different sets was analysed. This revealed no significant difference between them, with eukaryotes possessing about five TM 21-residue segments, compared with five 20-residue TM segments for prokaryotes.

The lower accuracy for eukaryotic proteins has several potential explanations. Eukaryotic proteins have an intrinsically more complex and variable range of topologies and thus are harder to predict. Alternatively, prokaryotic proteins may have a less comprehensive representation of the total proteins *in vivo*. The first explanation was tested by comparing the average number and length of TM regions for eukaryotic and prokaryotic proteins in the data-set: no significant difference was found. When the numbers of species contributing to each set was scrutinised, there initially appeared to be little evidence for the second hypothesis: of the 86 eukaryotic proteins, there are 20 different species represented, compared with 29 for the 130 prokaryotic proteins. However, analysis of individual species contribution reveals a different picture: 55% of the prokaryotic proteins were obtained from *Escherichia coli*. This is unsurprising, as *E. coli* is probably the most comprehensively studied of all prokaryotes, owing to its use in biological research. By contrast, the eukaryotic set is more widely spread across species, but is still dominated by 4 organisms, from which 75% of the proteins are derived: human (27%), cattle (24%), rat (15%) and yeast (10%). The biased prokaryotic data-set may increase the accuracy of prediction, as the method would be skewed significantly towards *E. coli* protein topology prediction. As these proteins obviously constitute the same large proportion of both test- and training-sets, then the predictor would predict them with improved accuracy.

Sequence typ	LOO cross- validation (%	5-fold cross-validation (%	MCC (%
Prokaryotic	77.4	75.2	0.856
Eukaryotic	61.4	58.6	0.795

Table 1: Performance of the alpha-helical predictor

The small range of organisms with well-characterised TM proteins is, as always, a significant challenge for predictor development. A comprehensive, but non-redundant, data-set is required if reliable and comprehensive predictive methods are to be developed. A predictor trained on the widest possible range of topologies is more likely to make accurate or, at least, more unbiased predictions. In the future it will be

ISSN 0973-2063
Bioinformation 1(6): 234-236 (2006)

necessary to continually retrain the α-helical predictor using newly published topologies, as they become available. This will hopefully increase the representation of different types of protein from all organisms.

Conclusion:

The method described here represents an important advance in the computational prediction of membrane protein structural class and topology. The α-helical TM protein topology predictor is of comparable accuracy to the best publicly available methods. The method described offers a useful alternative, yet 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] D. Frishman & H. W. Mewes, *Nat. Struct. Biol.,* 4:626 (1997) [PMID: 9253410]

[06] G. von Heijne, *Bioessays,* 17:25 (1995) [PMID: 7702590]

[07] J. Deisenhofer, *et al.*, *Methods Enzymol.,* 115:303 (1985) [PMID: 4079791]

[08] M. B. Ulmschneider & M. S. Sansom, *Biochim. Biophys. Acta.,* 1512:1 (2001) [PMID: 11334619]

[09] S. Moller, *et al., Bioinformatics,* 17:646 (2001) [PMID: 11448883]

[10] G. von Heijne, *Biochim. Biophys. Acta.,* 947:307 (1988) [PMID: 3285892]

[11] B. Traxler, *et al., J Membr. Biol.,* 132:1 (1993) [PMID: 8459445]

[12] M. L. Jennings, *Annu. Rev. Biochem.,* 58:999 (1989) [PMID: 2673027]

[13] B. Hille, Ion channels of excitable membranes. (2001), Sunderland, Mass.; [Great Britain]: Sinauer Associates

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

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

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

Edited by P. Kangueane

Citation: Taylor *et al.*, Bioinformation 1(6): 234-236 (2006)