Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Prediction Model**

A predictor of membrane class: Discriminating α-helical and β-barrel membrane proteins from non-membranous proteins

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 September 20, 2006; accepted October 02, 2006; published online October 07, 2006

Abstract:

Accurate protein structure prediction remains an active objective of research in bioinformatics. Membrane proteins comprise approximately 20% of most genomes. They are, however, poorly tractable targets of experimental structure determination. Their analysis using bioinformatics thus makes an important contribution to their on-going study. Using a method based on Bayesian Networks, which provides a flexible and powerful framework for statistical inference, we have addressed the alignment-free discrimination of membrane from non-membrane proteins. The method successfully identifies prokaryotic and eukaryotic α-helical membrane proteins at 94.4% accuracy, β-barrel proteins at 72.4% accuracy, and distinguishes assorted non-membranous proteins with 85.9% accuracy. The method here is an important potential advance in the computational analysis of membrane protein structure. It represents a useful tool for the characterisation of membrane proteins with a wide variety of potential applications.

Keywords: α-helical membrane proteins; β-barrel membrane proteins; membrane protein discrimination; Bayesian Network; alignment-free prediction

Background:

Accurate and reliable prediction of protein structure has long been a principal challenge for bioinformatics. Of particular importance is the prediction of membrane protein structure, as, unlike soluble and fibrous proteins, membrane proteins remain poorly tractable targets for the main experimental methods of structure determination: X-ray crystallography and multidimensional nuclear magnetic resonance (NMR) spectroscopy. **[1]** The seriousness of 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 Research Collaboration for Structural Bioinformatics (RCSB) Protein Data Bank (PDB). **[3, 4]**

Membrane proteins fall into two structural classes: α-helical and β-barrel. α-helical membrane proteins are responsible for interactions between most cells and their environment. Transmembrane (TM) helices are typically encoded by stretches of 17-25 residues, which provide sufficient length to cross the membrane. **[5]** A compositional bias towards hydrophobic residues is apparent in the TM helices, as they must make complementary interactions with the hydrophobic lipid bilayer. 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. **[6]** At present, the only known location for TM β-barrels is the outer membrane of Gram-negative bacteria. The SCOP database classifies TM β-barrels into 6 structural superfamiles: OmpA-like, OmpT-like, OmpLA, porins, TolC and Leukocidin (α Haemolysin). **[7]**

When considering algorithms that target problems in membrane proteomics, relatively few address the issue of distinguishing alpha helical transmembrane, beta transmembrane, and non-membraneous proteins. HUNTER **[8]**, for example, specifically addresses this issue: the algorithm has been tested on Gram-negative genomes with good accuracy for well- and partially-annotated proteins.

This paper describes an alignment-free prediction methodology that can distinguish between membrane and non-membrane proteins. These methods are based on Bayesian Networks (BNs), a form of machine learning that has been used very successfully in a number of biological applications in recent years. **[9, 10]** 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. **[11]**

Methodology:

Data-set

In compiling the membrane-class predictor data-set, the only requirement was for proteins of known sub-cellular location rather than accurate topologies. **[12]** The β-barrel set was smaller than the α-helical set owing to the lack of solved structures. The β-barrel set was taken from the PSORT-B complete data-set, version 1.1. **[13]** Non-membranous sequences were extracted from the Reinhardt and Hubbard data-set 14. **[14]** The number of sequences used from each compartment is listed in Table1.

Bayesian Network construction

A static full Bayesian model was used since such a model, compared with a naïve network model, outputs a probability that is not a product of probabilities from each descriptor but rather associates one probability with combinations of descriptors. Thus, overall performance is at least as good as that of the best individual descriptor.

We define the output prediction node of the network as O, the individual scale nodes are defined as S_1, S_2, *etc.*, individual scale node values are designated x_1, x_2, *etc.* The network models the joint probability distribution of all individual descriptor nodes. Predictions are made using Eqn 1:

$$P(O, S_1, \ldots, S_{434}) = P(O \mid S_1, \ldots, S_{434}) \prod_{i=1,434} P(S_i). \quad \text{Eqn 1}$$

Parameters of the network (probability tables) are defined thus:

$$A(x_1, \ldots, x_{434}, y) = P(S_1 = x_1, \ldots, S_{434} = x_{434}, O = y). \quad \text{Eqn 2}$$

$$B_k(x_1) = P(S_k = x_i), k = 1, \ldots, 434. \quad \text{Eqn 3}$$

Maximum likelihood estimation without priors is used, and therefore, if C is the empirical frequency of the parameter in the data and r is the amino acid being considered, a probability table for a range of scale node states:

$$A(x_1, \ldots, x_{434}, y) \sim \sum_r C(S_1^r, \ldots, S_{434}^r, O^r = y). \quad \text{Eqn 4}$$

Given this model, the optimal combined prediction is now defined as:

$$\arg\max_o P(o \mid S_1, \ldots, S_{434}) = \arg\max_o A(S_1, \ldots, S_{434}, o). \quad \text{Eqn 5}$$

Membrane protein class prediction method

The membrane class predictor seeks to classify membrane proteins and their structural class. Three classifications can be made: α-helical membrane protein, β-barrel membrane protein and non-membranous protein. Accordingly, the network is trained on α-helical, β-barrel and non-membranous sequences. Instead of attempting to classify local regions, this algorithm considers the whole protein using amino acid pseudo-composition. Pseudo-amino acid composition is used in preference to simple amino acid composition, as it attempts to model sequence-order effects, and hence more information about the sequence is used. Exploiting such additional information may prove useful, as proteins show different residue preferences in different parts of the sequence.

Pseudo-amino acid composition

Consider a protein of L residues:

$$R_1 R_2 R_3 \cdots R_L. \quad \text{Eqn 6}$$

The sequence-order effects can be approximated by sequence order-correlated factors:

$$\begin{cases} \theta_1 = \dfrac{1}{L-1}\sum_{i=1}^{L-1}\Theta(R_i, R_{i+1}) \\ \theta_2 = \dfrac{1}{L-2}\sum_{i=1}^{L-2}\Theta(R_i, R_{i+2}) \\ \theta_3 = \dfrac{1}{L-3}\sum_{i=1}^{L-3}\Theta(R_i, R_{i+3}), \quad (\lambda < L) \\ \cdots \\ \theta_\lambda = \dfrac{1}{L-\lambda}\sum_{i=1}^{L-\lambda}\Theta(R_i, R_{i+\lambda}) \end{cases} \quad \text{Eqn 7}$$

θ_1 is the first-tier correlation factor reflecting the sequence-order correlation between adjacent residues in the sequence; θ_2 is the first-tier correlation factor reflecting correlation between residues two positions apart in the sequence; and so on as the sequence separation increases. $\Theta(R_i, R_j)$ is the correlation function, defined by:

$$\Theta(R_i, R_j) = \frac{1}{3}\left\{\left[H_1(R_j) - H_1(R_i)\right]^2 + \left[H_2(R_j) - H_2(R_i)\right]^2 + \left[M(R_j) - M(R_i)\right]^2\right\}, \text{ Eqn 8}$$

where $H_1(R_i)$, $H_2(R_i)$ and $M(R_i)$ represent the residue hydrophobicity, hydrophilicity and side-chain mass of residue R_j. These values are then normalised:

$$\begin{cases} H_1(i) = \dfrac{H_1^0(i) - \sum_{i=1}^{20} \frac{H_1^0(i)}{20}}{\sqrt{\sum_{i-1}^{20}\left[\frac{H_1^0(i) - \sum_{i=1}^{20}\frac{H_2^0(i)}{20}}{20}\right]^2}} \\ H_2(i) = \dfrac{H_2^0(i) - \sum_{i=1}^{20} \frac{H_2^0(i)}{20}}{\sqrt{\sum_{i-1}^{20}\left[\frac{H_1^0(i) - \sum_{i=1}^{20}\frac{H_2^0(i)}{20}}{20}\right]^2}} \\ M(i) = \dfrac{M^0(i) - \sum_{i=1}^{20} \frac{M^0(i)}{20}}{\sqrt{\sum_{i-1}^{20}\left[\frac{M^0(i) - \sum_{i=1}^{20}\frac{M^0(i)}{20}}{20}\right]^2}} \end{cases} \quad \text{Eqn 9}$$

$H_1^0(i)$ is the hydrophobicity value of the *i*th residue,as defined by Tanford [64]; $H_2^0(i)$ is the hydrophilicity values of the *i*th residue, as defined by Hoop and Woods [65]; and $M^0(i)$ is the mass if the *i*th amino acid side-chain.

Pseudo-amino acid composition is calculated using a formula that assigns 40 scores for each sequence X, 20 representing normal amino acid composition (x_1, …, x_{20}) and 20 representing sequence-order effects (x_{20+1}, …, $x_{20+\lambda}$):

$$X = \begin{bmatrix} x_1 \\ \bullet \\ \bullet \\ \bullet \\ x_{20} \\ x_{20+1} \\ \bullet \\ \bullet \\ \bullet \\ x_{20+\lambda} \end{bmatrix}, \qquad \text{Eqn 10}$$

where

$$x_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{20} f_i + w\sum_{j=1}^{\lambda}\theta_j}, & (1 \le u \le 20) \\ \dfrac{w\theta_{u-20}}{\sum_{i=1}^{20} f_i + w\sum_{j=1}^{\lambda}\theta_j}, & (20+1 \le u \le 20+\lambda) \end{cases}, \qquad \text{Eqn 11}$$

where f_i represents the normalised frequency of the 20 residues in X, θ_j represents the *j*-tier sequence-correlation factor for X, and *w* is the weight factor for the sequence-order effect, which was set at 0.05.

Assessing Prediction Accuracies

Accuracies were assessed using two methods of cross-validation: leave-one-out (LOO) and five-fold cross-validation. LOO cross-validation removes one protein from the data-set, trains the network on the remaining proteins and then tests the removed protein. The test is repeated, removing and testing different proteins, until all proteins have been tested. Five-fold cross-validation randomly extracts 1/5 of the data-set and then retrains the network on the remaining 4/5 and tests the removed 1/5. This is repeated 4 times, each iteration excluding proteins that were previously included; therefore, each individual protein is included in the test set only once.

Protein class	Number of sequences
Eukaryotic membranous	432
Eukaryotic non-membranous	2106
Prokaryotic inner membrane	268
Prokaryotic outer membrane	352
Prokaryotic non-membranous	997

Table 1: Numbers of sequences used in the protein-class predictor

	LOO cross-validation	Five-fold cross-validation	MCC
All classes	85.2	82.4	0.915
α -helical	94.4	96.5	0.932
α -barrel	72.4	72.1	0.835
Non-membranous	85.9	84.6	0.868

Table 2: Performance of the protein-class predictor when trained on pseudo- composition

	LOO cross-validation	Five-fold cross-validation	MCC
All classes	74.84	75.72	0.805
α -helical	88.68	85.89	0.896
α -barrel	48.97	51.23	0.629
Non-membranous	81.68	79.99	0.816

Table 3: Performance of the protein-class predictor when trained on simple amino-acid composition

Results and Discussion:

A BN was constructed and then trained on membrane proteins, both α-helical and β-barrel, and non-membranous proteins for a variety of sub-cellular compartments. The sequences (with the obvious exception of the β-barrels) were of mixed eukaryotic and prokaryotic origin. As shown in Table 2, the results are of good accuracy, α-helical proteins being the most successfully identified (94.4%/96.5%), β-barrel proteins proving harder to classify, with accuracies of 72.4%/72.1%. To test whether amino acid pseudo-composition **[15]** provides a significant increase in accuracy from a BN trained on simple amino acid composition, we also trained networks using amino acid proportions only. See Table 3. These results indicated that networks based on pseudo-composition were significantly more accurate classifiers.

The most significant difference between pseudo and simple composition is in the Matthews Correlation Coefficient (MCC), which decreases by 0.110 when predicting "all classes". This change is mostly the result of an increased rate of false-positive detection. A 0.036 fall in α-helical protein predictions was observed, and a 0.186 fall in β-barrel MCC owing to increased false-positives. Thus results for the protein-class predictor display considerable differences between the accuracies for the different locations. There are several possible explanations for this variance. The method relies on the use of pseudo-amino acid composition to discriminate compartments. It is thus probable that sequences that mimic the composition of membrane proteins will be predicted as membranous. This is a problem particularly common in proteins that are secreted, as has been observed by other researchers, and is often attributed to the N-terminal signal sequence. **[16]** The signal sequence has a hydrophobic core and averages 20-30 residues in length, making it highly similar to a TM α-helix, which often causes problems for α-helical membrane protein topology predictors. **[17]** This problem is best addressed by using specific signal sequence prediction methods, which have high degrees of accuracy. **[18]** Other false predictions were found to result from stretches of hydrophobic residues that fold inside globular proteins, but this was only observed in 2 cases. False-positive α-helix predictions are much less prevalent using machine-learning methods, such as described here, than when using simple hydrophobic plots, but still represent an area for future improvement.

Of special note is the observation that no false-positive β-barrel predictions were made. As β-barrel proteins often have varied compositions, being exposed to a range of environments in different segments of the proteins, one might assume that they would likely possess amino acid characteristics that partially imitate other sub-cellular compartments or have no overall characteristic composition. Our results show that the opposite is true. This may indicate that the exposure of β-barrels to a range of environments produces a more characteristic amino acid composition than expected. To further investigate this, we examined the simple amino acid compositions of all proteins used. As expected, a skewed bias towards hydrophobic amino acids is observed in the α-helical membrane proteins; the non-membranous proteins present a generalised composition, as expected from their different locations and functions; and the β-barrel composition differs from both, but shows no simple, overall pattern of preference. This suggests that this apparent complexity is captured well by pseudo-composition, giving the network its predictive power.

To assess the ability of the protein-class predictor to aid genome annotation, the predictor was used to classify all proteins of known sub-cellular location from both the human and *E. coli* genomes. The proteins were obtained from Swiss-Prot release 42, and only proteins of unambiguous location were used. From the human genome, 5568 proteins were tested, of which 2416 were membranous. The protein-class predictor correctly identified 90.54% of the membrane proteins and 82.58% of the non-membranous proteins. For *E. coli,* the results were of higher accuracy, 97.68% of the 689 membranous proteins being correctly classified, and 91.89% of the non-membranous proteins. These results indicate that the protein-class predictor is a powerful tool for genome annotation, able to differentiate protein class with a high level of confidence. The higher rate of accuracy for *E. coli* proteins probably reflects their over-representation in the training-set.

Conclusion:

The method described here represents an important advance in the computational determination of membrane protein structural class and topology. It provides an accurate tool for the alignment-free classification of proteins into TM α-helical, TM β-barrel or non-TM. Although many predictors can distinguish between α-helical membrane proteins and other proteins, few have been reported that can reliably predict β-barrel membrane proteins, and fewer still have been produced that combines both functions. The protein-class predictor provides both a means of annotating the location of novel proteins and an *in silico* tool to aid the discovery of drug and vaccine targets. Thus, the method offers a useful approach 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] J. Deisenhofer, *et al*., *Methods Enzymol.*, 115:303 (1985) [PMID: 4079791]
[06] S. Moller, *et al*., *Bioinformatics*, 17:646 (2001) [PMID: 11448883]
[07] A. G Murzin, *et al*., *J. Mol. Biol.*, 247:536 (1995) [PMID: 7723011]
[08] R. Casadio, *et al*., *Prot. Sci.*, 12:1158 (2003) [PMID: 12761386]
[09] V. Pavlovic, *et al*., *Bioinformatics*, 18:19 (2002) [PMID: 11836207]
[10] C. S. Schmidler, *et al*., *J. Comput. Biol.*, 7:233 (2000) [PMID: 10890399]
[11] J. Pearl, *Probablistic Reasoning in Intellegent Systems: Networks of Plausible Inference*. San Mateo, California: Morgan Kaufman, (1988)
[12] M. Ikeda, *et al*., *Nucleic Acids Res.*, 31:406 (2003) [PMID: 12520035]
[13] J. L. Gardy, *et al*., *Nucleic Acids Res.*, 31:3613 (2003) [PMID: 12824378]
[14] A. Reinhardt & T. Hubbard, *Nucleic Acids Res*., 26:2230 (1998) [PMID: 9547285]
[15] K. C. Chou, *Proteins*, 43:246 (2001) [PMID: 11288174]
[16] H. Nielsen, *et al*., *Protein Eng.*, 12:3 (1999) [PMID: 10065704]
[17] J. Nilsson, *et al*., *FEBS Lett.*, 486**:**267 (2000) [PMID: 11119716]
[18] J. D. Bendtsen, *et al*., *J. Mol. Biol.*, 340:783 (2004) [PMID: 15223320]

Edited by P. Kangueaane
Citation: Taylor *et al*., Bioinformation 1(6): 208-213 (2006)