WSEAS TRANSACTIONS on BIOLOGY and BIOMEDICINE

Issue 2, Volume 1, April 2004

ISSN 1109-9518 http://www.wseas.org

EDITORIAL BOARD

TOPICS: Molecular Dynamics, Biochemistry, Biophysics, Quantum Chemistry, Molecular Biology, Cell Biology, Immunology, Neurophysiology, Genetics, Population Dynamics, Dynamics of Diseases, Bioecology, Epidemiology, Photobiology, Photochemistry, Plant Biology, Bioinformatics, Signal Transduction, Environmental Systems, Psychological and Cognitive Systems, Evolution, Game Theory and Adaptive Dynamics, Bioengineering, Biotechnologies, Medical Imaging, 2-dimensional and 3-Dimensional Signal Processing, Nuclear Biology and Medicine, Cancer Therapy and Mathematical Models.

HOW TO SUBMIT: http://www.wseas.org, http://www.worldses.org

SUBSCRIPTION: The subscription rate for each journal is 100 Euros (per year) for individuals and 200 Euros (per year) for institutions or companies.

FORMAT OF THE PAPERS: http://www.worldses.org/journals

ISSN: 1109-9518

USA Office
Prof. George Antoniou,
Montclair State University,
Computer Science
Department,
New Jersey, USA.

http://www.worldses.org

European Office I
WSEAS
Ag. I. Theologou 17-23,
15773, Zographou,
Athens, GREECE.
Tel: (+30) 210 7473313
Fax: (+30) 210 7473314
http://www.wseas.org

European Office II
Prof. Nikos Mastorakis,
Military Institutes of Univ. Education,
Hellenic Naval Academy,
Department of Computer Science,
Terma Hatzikyriakou, 18539,
Piraeus, GREECE.
http://www.wseas.org

WSEAS E-LIBRARY: http://www.wseas.org/data
WSEAS CHAPTERS: http://www.wseas.org/chapters

Each paper of this issue was published after review by 3 independent reviewers

Finding beta-barrel outer membrane proteins with a Markov Chain Model

PANTELIS G. BAGOS[1], THEODORE D. LIAKOPOULOS[2]
and STAVROS J. HAMODRAKAS[3]
[1,2,3]Department of Cell Biology and Biophysics, Faculty of Biology
University of Athens
Panepistimiopolis, Athens 15701
GREECE
[1]pbagos@biol.uoa.gr
[2]liakop@biol.uoa.gr
[3]shamodr@cc.uoa.gr
http://bioinformatics.biol.uoa.gr/mcmbb

Abstract: - The task of finding β-barrel outer membrane proteins of the gram-negative bacteria is of great importance in current Bioinformatics research. We developed a computational method, which discriminates β-barrel outer membrane proteins from globular ones and, also, from α-helical membrane proteins. The method is based on a 1st order Markov Chain model, which captures the alternating pattern of hydrophilic-hydrophobic residues occurring in the membrane-spanning beta-strands of beta-barrel outer membrane proteins. The model achieves high accuracy in discriminating outer membrane proteins, and could be used alone, or in conjunction with other more sophisticated methods, already available.

Key-Words: - outer membrane proteins, beta barrel, Markov chains, bioinformatics.

1 Introduction

β-barrel membrane proteins, constitute one of the two main structural classes of integral membrane proteins. They are located in the outer membrane of gram-negative bacteria, and presumably in the outer membrane of chloroplasts and mitochondria. These proteins have their membrane spanning segments formed by antiparallel β-strands, creating a channel in the form of a barrel that spans the outer membrane [1]. This is in contrast to the α-helical membrane proteins of the cytoplasmic membrane of all cells, that have their membrane spanning regions forming α-helices, which mainly consist of hydrophobic residues [2]. Whereas the prediction of transmembrane regions and consequently the genome-wide prediction of α-helical membrane proteins is nowadays a relatively easy task, this is not the case for the β-barrel membrane proteins. This is due to the lack of a clear pattern in their membrane spanning strands, such as the stretch of 15-30 consecutive hydrophobic residues or the Positive Inside rule, which occur in the α-helical proteins. Furthermore, discrimination of transmembrane strands from other β-strands, forming β-barrel structures in water-soluble proteins, is even more difficult. The reason for that is the fact that water-soluble proteins that form β-barrel structures, share (up to a certain degree) common features with the transmembrane strands of the bacterial outer membrane proteins, such as amphipathicity. The β-barrel outer membrane proteins perform a wide variety of functions such as active ion transport, passive nutrient uptake, membrane anchoring, adhesion, and catalytic activity. A large number of pathogens are actually bacteria belonging to the gram-negative bacteria class, and for those bacteria the virulence activity in a lot of cases has been proven to depend on specific outer membrane proteins. Considering additionally the important biological functions in which outer membrane proteins are involved in, it is not a surprise that these proteins attract an increased medical interest. This is confirmed by the continuously increasing number of completely sequenced genomes of gram-negative bacteria deposited in the public databases. A few approaches have been made, in the direction of predicting the transmembrane strands of outer membrane proteins and/or identifying those proteins when searching large data sets; they are based on studies of the physico-chemical properties of the β-strands, such as hydrophobicity and amphipathicity [3], statistical analyses based on the amino acid composition of the known structures [4], or machine learning techniques like neural network predictors [5], and

Hidden Markov Models [6, 7, 8]. Recently, two methods based on HMMs [7, 8], achieved the highest accuracy.

In this work, we developed a computationally rather simple and fast method that discriminates with high accuracy and precision β-barrel outer membrane proteins in large datasets. The method is based on a Markov Chain model [9], which captures the alteration of hydrophilic-hydrophobic residues in the transmembrane β-strands of outer membrane proteins, while, at the same time it does not predict a large number of false positives. The model was trained on a non-redundant set of 121 experimentally verified outer membrane proteins, and has been tested with a jackknife procedure, yielding 89.26% and 92.67% correct classification rate for outer membrane and globular proteins, respectively. Furthermore, the model produces no false positive results when screening a dataset of 276 experimentally verified α-helical membrane proteins set.

2 Materials and Methods

In sub-section 2.1 we present the algorithmic details, of the Markov chain model we used, whereas in sub-section 2.2 we describe the datasets used for training and testing the method. In this section we are mainly using the notation of Durbin *et al.* [9].

2.1 The Markov Chain.

If we denote an amino-acid sequence of length *n*, by **x** such as:

$$\mathbf{x} = x_1, x_2, \dots, x_{n-1}, x_n$$

and consider the amino-acid distribution at each position *i* along the sequence as a random variable, then we can define a Markov chain as a stochastic process with what is called a Markov Property. In the discrete case (such as in our case), the process consists of the sequence *x* of random variables taking values in a "state space" defined on the alphabet of the amino-acids, the value of x_i being "the state of the system at time *i*". The (discrete-time) Markov property states that the conditional distribution of the "future"

$x_{i+1}, x_{i+2}, x_{i+3}, \dots$ given the "past", $x_1, x_2, \dots, x_{i-1}, x_i$, depends on the past only through x_i. In other words, knowledge of the most recent past state of the system renders knowledge of less recent history irrelevant. This is formulated by:

$$P(x_i \mid x_{i-1}, \dots, x_1) = P(x_i \mid x_{i-1}) \qquad (1)$$

Each particular Markov chain may be identified with its matrix of "transition probabilities", often called simply its transition matrix [9]. The entries in the transition matrix are given by:

$$a_{st} = P(x_i = t \mid x_{i-1} = s) = \alpha_{x_{i-1}x_i}$$

and this is the probability of residue *t* occurring at position *i* in the sequence, given that the preceding residue (*i-1*) is *s*. Considering that we can generalize the dependence over *k* "past" (preceding) residues, this kind of Markov Chain is usually denoted a 1st order Markov Chain. The total probability of a sequence is computed according to:

$$P(\mathbf{x}) = P(x_1, x_2, \dots, x_{n-1}, x_n) =$$
$$= P(x_n \mid x_{n-1}, \dots, x_1) P(x_{n-1} \mid x_{n-2}, \dots, x_1) \dots P(x_1)$$

and from (1) we have:

$$P(\mathbf{x}) = P(x_n \mid x_{n-1}) P(x_{n-1} \mid x_{n-2}) \dots P(x_2 \mid x_1) P(x_1)$$
$$= P(x_1) \prod_{i=2}^{n} P(x_i \mid x_{i-1}) = P(x_1) \prod_{i=2}^{n} \alpha_{x_{i-1}x_i}$$

where $P(x_i)$ is the probability for the starting symbol. The Maximum Likelihood Estimates (MLEs) of the transition probabilities [9], are computed according to:

$$\hat{\alpha}_{x_{i-1}x_i} = \frac{c_{x_{i-1}x_i}}{\sum_{x_i'} c_{x_{i-1}x_i'}}$$

where c_{st} are the observed counts of residue *s* followed by residue *t* in the training sequences, and the sum in the denominator extends over the hole alphabet of the 20 amino-acids. Assuming two different models, using different transition probabilities matrices (a model for the positive examples denoted by +, and a model for the negative examples denoted by –), we can define a log-odds score *S(x)* for the entire sequence, which is useful for discrimination purposes:

$$S(\mathbf{x}) = \log \frac{P(\mathbf{x} \mid +)}{P(\mathbf{x} \mid -)} = \sum_{i=1}^{n} \log \left(\frac{\alpha^{+}_{x_{i-1}x_i}}{\alpha^{-}_{x_{i-1}x_i}} \right) = \sum_{i=1}^{n} \beta_{x_{i-1}x_i}$$

where $\beta_{x_{i-1}x_i}$, is the log-odds for the transition from residue x_{i-1} to x_i and it is a measure of the propensity for that transition probability to occur more frequently in to one or the other model. Values of $\beta_{x_{i-1}x_i}$ larger than zero indicate preference for model +, whereas values smaller than zero a preference for model -. In order to eliminate the influence of the sequence length on the total score, we further normalize by dividing with the length *n* of the sequence, thus producing a normalized score (per residue log-odds score).

$$S^{norm}(\mathbf{x}) = \frac{S(\mathbf{x})}{n} = \frac{\sum_{i=1}^{n} \beta_{x_{i-1}x_i}}{n} \qquad (2)$$

Considering higher-order (k^{th}) Markov Chains, the generalization of Equation (1) is straightforward to include the dependence on k residues back in the sequence:

$$P(x_i \mid x_{i-1},\ldots,x_1) =$$
$$= P(x_i \mid x_{i-1}, x_{i-2},\ldots,x_{i-k}) = \alpha_{x_k \ldots x_{i-1}x_i}$$

since

$$P(x_i \mid x_{i-1}, x_{i-2},\ldots,x_{i-k}) =$$
$$= P(x_i, x_{i-1},\ldots,x_{i-k+1} \mid x_{i-1}, x_{i-2},\ldots,x_{i-k})$$

the k^{th} order Markov Chain reduces to a 1st order one, over an alphabet of size 20^k, thus requiring the calculation of a transition matrix of 20^{k+1}x20^{k+1} transition probabilities. Thus, whereas for a 1st order model we needed to calculate 20^2=400 transition probabilities, for a 3rd order model we need 20^3=8000, and when the number of sequences used for training is limited, this could lead to over-fitting and to an inadequate training. In general, the higher the order of the model the better would be the discriminative power, but practical limitations arising from the size of the sequence database used for training, forced us to be parsimonious concerning the order of the model.

2.2 Training and testing sets

For training the model we compiled a non-redundant dataset of well-annotated β-barrel outer membrane proteins. We collected the sequences belonging to the dataset used in the validation of the PSORT-B algorithm [10]. The sequences have been submitted to a redundancy check, removing chains with a sequence identity above some threshold. We considered two sequences as being homologues, if they demonstrated an identity above 30% in a pairwise alignment, in a length longer than 80 residues. For the pairwise local alignment we used BlastP [11] with default parameters, and the homologous sequences were removed implementing Algorithm 2 from Hobohm et al [12]. The remaining 121 outer membrane proteins constitute our positive examples training set.

As a negative examples set, we used an additional dataset of globular proteins, with known 3-dimensional structures deposited in PDB [13]. This set was compiled using the PAPIA [14] server, with the sequence similarity threshold set to 25%, and excluding membrane proteins, proteins with a length lower than 80 residues, and proteins with at least one unidentifiable residue in the sequence; finally we came up with 1133 sequences of such globular proteins.

To further test the ability of the model to correctly predict outer membrane β-barrel proteins, we used few additional sets. Thus, we used the entire TMPDB database [15], which contains 276 α-helical membrane proteins with experimentally determined topology, in order to examine the ability of the model to discriminate outer membrane β-barrel proteins from α-helical ones. As an independent test set of β-barrel outer membrane proteins, we used the 149 sequences belonging to the sub-class β-barrel porins, of the class channels/pores of the TCDB [16].

As a final independent test set, we used a set of 100,000 simulated sequences with aminoacid composition similar to that of the Swiss-Prot database [17]. This was done in order to address the question regarding the rate of false positives occurring purely by chance.

3 Results

Using Equation (2) we were able to obtain a prediction for each protein, in the dataset. Simply, if *S(x)* is greater than zero the protein is predicted to be an outer membrane protein, otherwise it is not predicted to be an outer membrane protein. The accuracy of the trained model (self consistency) is rather high. It correctly predicts 112 out of the 121 outer membrane proteins in the training set (92.56%) and 1052 out of the 1100 globular proteins (92.85%). The model has also been tested with the well-known jacknife procedure which consists of removing a protein from the training set, training the model with the remaining proteins and performing the test on the protein removed. This process is tandemly repeated for all proteins in the training set, and the final prediction accuracy summarizes the outcome of all independent tests. In this jackknife test, the model correctly predicts 109 out of the 121 outer membrane proteins (90.08%) and 1050 out of the 1133 globular proteins (92.67%). These results are similar with those obtained by the HMMs in [7, 8].

In the independent set of 149 β-barrel membrane proteins from TCDB used for evaluation, the model correctly predicts 137 proteins (91.96%), whereas in the independent set of 276 α-helical membrane proteins derived from TMPDB, the model does not produce even one false positive (100%). Furthermore, in the 100,000 simulated sequences the model produces false positive results with an extremely low rate of 0.57%. All the above indicate that the Markov Chain model captures some of the special features of the β-barrel outer membrane proteins, and thus it could reliably used for predicting the nature of newly determined not-annotated proteins.

4 Conclusions

We presented here a Markov Chain model that discriminates with high accuracy and precision β-barrel outer membrane proteins. The model reaches similar predicting performance in the discrimination procedure with other more sophisticated methods such as the Hidden Markov Model, and it could be used in conjunction with them in order to achieve better predictions. The method is computationally simple, thus it is very fast and suitable for screening large datasets such as entire proteomes of gram-negative bacteria, in order to find novel outer membrane proteins. A web server running the application is available at the url: http://bioinformatics.biol.uoa.gr/mcmbb, where the user may submit up to 500 sequences and receive the prediction.

References:

[1] Schulz GE, The structure of bacterial outer membrane proteins. *Biochim Biophys Acta*, Vol. 1565, No. 2, 2002, pp. 308-17

[2] Von Heijne G, Recent advances in the outstanding of membrane protein assembly and function. *Quart Rev Biophys*, Vol.32, No 4, 1999, pp. 285-307.

[3] Zhai Y and Saier MH Jr, The beta-barrel finder (BBF) program, allowing identification of outer membrane beta-barrel proteins encoded within prokaryotic genomes. *Protein Sci*, Vol. 11, No. 9, 2002, pp. 2196-207.

[4] Wimley WC, Toward genomic identification of beta-barrel membrane proteins: composition and architecture of known structures. *Protein Sci*, Vol. 11, No. 2, 2002, pp. 301-12.

[5] Jacoboni I, Martelli PL, Fariselli P, De Pinto V and Casadio R, Prediction of the transmembrane regions of beta-barrel membrane proteins with a neural network-based predictor. *Protein Sci*, Vol. 10, No. 4, 2001, pp. 779-87.

[6] Liu Q, Zhu YS, Wang BH and Li YX, A HMM-based method to predict the transmembrane regions of beta-barrel membrane proteins. *Comput Biol Chem*, Vol. 27, No. 1, 2003, pp. 69-76.

[7] Martelli PL, Fariselli P, Krogh A and Casadio R, A sequence-profile-based HMM for predicting and discriminating beta barrel membrane proteins. *Bioinformatics*, Vol. 18, Suppl 1, 2002, pp. S46-53.

[8] Bagos PG, Liakopoulos TD, Spyropoulos IC and Hamodrakas SJ. A Hidden Markov Model capable of predicting and discriminating β-barrel outer membrane proteins. *BMC Bioinformatics*, Vol. 5, No. 29, 2004.

[9] Durbin R, Eddy S, Krogh A and Mithison G, *Biological sequence analysis, probabilistic models of proteins and nucleic acids*, Cambridge University Press, 1998.

[10] Gardy JL, Spencer C, Wang K, Ester M, Tusnady GE, Simon I, Hua S, deFays K, Lambert C, Nakai K and Brinkman FS, PSORT-B: Improving protein subcellular localization prediction for Gram-negative bacteria. *Nucleic Acids Res*, Vol. 31, No. 13, 2003, pp. 3613-7.

[11] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, Vol. 25, No. 17, 1997, pp. 3389-402.

[12] Hobohm U, Scharf M, Schneider R and Sander C, Selection of representative protein data sets. *Protein Sci*, Vol. 1, No. 3, 1992, pp. 409-17.

[13] Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, Feng Z, Gilliland GL, Iype L, Jain S, Fagan P, Marvin J, Padilla D, Ravichandran V, Schneider B, Thanki N, Weissig H, Westbrook JD and Zardecki C, The Protein Data Bank. *Acta Crystallogr D Biol Crystallogr*, Vol. 58, No 1, 2002, pp. 899-907.

[14] Noguchi T and Akiyama Y, PDB-REPRDB: a database of representative protein chains from the Protein Data Bank (PDB) in 2003. Nucleic Acids Res, Vol. 31, No. 1, 2003, pp. 492-3.

[15] Ikeda M, Arai M, Okuno T and Shimizu T, TMPDB, a database of experimentally-characterized transmembrane topologies, *Nucleic Acids Res*, Vol. 31, No. 1, 2003, pp. 406–409.

[16] Busch W and Saier MH Jr, The transporter classification (TC) system, 2002, *Crit. Rev. Biochem. Mol. Biol.*, Vol. 37, No. 5, 2002, 287-337.

[17] Boeckmann B, Bairoch A, Apweiler R, Blatter M-C, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, and Schneider M, The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003, *Nucleic Acids Res*, Vol 31, No. 1, 2003, pp. 365-370.