BIOINFORMATICS

Vol. 18 Suppl. 1 2002
Pages S46–S53

A sequence-profile-based HMM for predicting and discriminating β barrel membrane proteins

Pier Luigi Martelli[1], *Piero Fariselli*[1], *Anders Krogh*[2,3] *and Rita Casadio*[1,*]

[1]*Laboratory of Biocomputing, CIRB/Department of Biology, University of Bologna, via Irnerio 42, 40126 Bologna, Italy and* [2]*Center for Biological Sequence Analysis, the Technical University of Denmark, DK-2800 Lyngby, Denmark*

Received on January 24, 2002; revised and accepted on March 26, 2002

ABSTRACT

Motivation: Membrane proteins are an abundant and functionally relevant subset of proteins that putatively include from about 15 up to 30% of the proteome of organisms fully sequenced. These estimates are mainly computed on the basis of sequence comparison and membrane protein prediction. It is therefore urgent to develop methods capable of selecting membrane proteins especially in the case of outer membrane proteins, barely taken into consideration when proteome wide analysis is performed. This will also help protein annotation when no homologous sequence is found in the database. Outer membrane proteins solved so far at atomic resolution interact with the external membrane of bacteria with a characteristic β barrel structure comprising different even numbers of β strands (β barrel membrane proteins). In this they differ from the membrane proteins of the cytoplasmic membrane endowed with alpha helix bundles (all alpha membrane proteins) and need specialised predictors.

Results: We develop a HMM model, which can predict the topology of β barrel membrane proteins using, as input, evolutionary information. The model is cyclic with 6 types of states: two for the β strand transmembrane core, one for the β strand cap on either side of the membrane, one for the inner loop, one for the outer loop and one for the globular domain state in the middle of each loop. The development of a specific input for HMM based on multiple sequence alignment is novel. The accuracy per residue of the model is 83% when a jack knife procedure is adopted. With a model optimisation method using a dynamic programming algorithm seven topological models out of the twelve proteins included in the testing set are also correctly predicted. When used as a discriminator, the model is rather selective. At a fixed probability value, it retains 84% of a non-redundant set comprising 145 sequences of well-annotated outer membrane proteins. Concomitantly, it correctly rejects 90% of a set of globular proteins including about 1200 chains with low sequence identity (<30%) and 90% of a set of all alpha membrane proteins, including 188 chains.

Availability: The program will be available on request from the authors.

Contact: gigi@lipid.biocomp.unibo.it; http://www.biocomp.unibo.it

Keywords: β barrel membrane proteins; HMM; genome analysis; protein structure prediction; outer membrane protein topology.

INTRODUCTION

Membrane proteins have been known in the outer membrane of bacteria, chloroplasts and mitochondria for about ten years (Schulz, 2000). These chains, referred to as β-barrel membrane proteins (Schulz, 2000), comprise the archetypal trimeric porins of Gram-negative bacteria consisting of water-filled channels that non-specifically mediate the passive transport of ions and small hydrophilic molecules (<6KD) or select for certain molecules such as malto-oligosaccharides. More recently, other β-barrel membrane proteins have been characterized with quite diverse functions from that of archetypal porins, especially in enteric bacteria. Presently it is known that the β-barrel structure is associated to functions that are more and more relevant to the entire cell metabolism and are as diverse as active ion transport, passive nutrient intake, membrane anchors, membrane-bound enzymes and defence against attack proteins. In addition, it is now evident that the different functions are associated to different barrel sizes (ranging from small eight-stranded to large twenty-two stranded β barrels), to different topologies and aggregation number (Schulz, 2000).

Although after a decade of analysis the construction principles of β-barrel membrane proteins are known (Schulz, 2000), it is almost impossible to derive three-

*To whom correspondence should be addressed.

[3]Present address: Bioinformatics Centre, University of Copenhagen, DK-2100 Copenhagen, Denmark

dimensional models for proteins of the outer membrane. This is due to the fact that unless they belong to the same family, β-barrel membrane proteins share little sequence identity even in the transmembrane spanning regions.

It is therefore necessary to be able to correctly locate the transmembrane regions in a sequence in order to assign the correct barrel topology and eventually build a three-dimensional model on the basis of the few existing templates, solved at atomic resolution.

Recently a neural-network-based predictor was developed, especially suited to predict the topography of beta barrel transmembrane proteins (Jacoboni *et al.*, 2001). This task however appears to be more difficult than predicting the topography and topology of all-helical membrane proteins, whose transmembrane domains can be fairly well detected both with statistical methods, neural networks and HMMs (Jones *et al.*, 1994; Rost *et al.*, 1995, 1996; Tusnady and Simon, 1998; Krogh *et al.*, 2001).

In this paper we use the prototypes of the β-barrel membrane proteins crystallized so far for training and testing a method based on HMM. The model is novel and is trained on evolutionary information as computed from sequence profile. The method trained and tested with a jack knife procedure on the 12 proteins included in the database reaches an overall accuracy per residue as high as 83%. In addition with a model optimisation method (based on a dynamic programming algorithm) seven topological models out of the 12 proteins are correctly predicted. When a large-scale sequence analysis is performed the model is also capable of correctly discriminating outer membrane from all helical membrane proteins and from globular proteins.

SYSTEM AND METHODS

Sequence-profile-based HMM

To implement the model based on HMM that analyses the evolutionary information encoded in a sequence profile, we replace the symbolic sequences of characters (that are currently analysed by standard HMMs (Durbin *et al.*, 1998)) with sequences of vectors. These contain the sequence profile computed from a multiple sequence alignment. If L is the length of the chain and A is the size of the alphabet over which vectors are built (that is $A = 20$ for proteins), we refer to this sequence vector with the notation:

$$\boldsymbol{s} = \boldsymbol{s}^1\boldsymbol{s}^2\ldots\boldsymbol{s}^L = (s^1(1), s^1(2), \ldots s^1(A))(s^2(1), s^2(2), \ldots s^2(A)) \ldots (s^L(1), s^L(2), \ldots s^L(A)) \quad (1)$$

The components of each vector $\boldsymbol{s}^t$ are positive and sum to a constant value S (independent of the position t):

A multiple sequence alignment based HMM is composed of a Markov model with N states connected by means of the transition probabilities a_{ij}. The probability density function for the emission of a vector from each state is determined by a number A of parameters that are peculiar for each state k and are indicated with the symbols $e_k(c)$ (with $c = 1, 2, \ldots, A$):

$$P(\boldsymbol{s}^t \mid \pi^t = k) = (1/Z) \cdot \sum_c s^t(c) \cdot e_k(c), \quad (2)$$

where π^t is the tth state in the path. Z is the normalising factor:

$$Z = \iint_{\text{constraints}} \mathrm{d}\boldsymbol{s}^t \left(\sum_c s^t(c) \cdot e_k(c) \right) = (S^A/A!) \times \sum_c e_k(c) \quad (3)$$

with $\sum_c e_k(c) = 1$.

In principle the value of Z depends both on the position t along the sequence and on the current state. It can be proven that if the emission parameters sum to a constant independent of the state then also the value of Z is independent of the state and of the position along the sequence (see Appendix 1).

Dynamic programming algorithms for sequence-profile-based HMM

The result in Equation 3 indicates that the normalisation constant of the emission probability of a vector is independent of the vector, and depends on the dimension of the vectors, A, and on the constant S that is the sum of all the elements of a vector. Under these conditions, irrespective of the value of Z, it is possible to apply standard techniques for estimation and decoding to the sequence-profile-based HMMs.

With the substitution of the emission probability $e_k(s^t)$ of a standard HMM with

$$(1/Z) \cdot \sum_c e_k(c) \cdot s^t(c) \quad (4)$$

all the standard algorithms are essentially the same (see e.g. (Durbin *et al.*, 1998)).

Since Z is a constant it can be omitted while computing the recursive variables. The contribution of Z to the probability values is considered only in the final iteration of the recursive procedure: each position of the sequence contributes by a factor $1/Z$.

EM algorithm for sequence-profile-based HMM

Training of sequence-profile-based HMM is based on a generalised form of the Expectation-Maximisation algorithm (Durbin *et al.*, 1998).

Updating of the transition and emission probabilities is performed by computing:

$$a_{ij} = A_{ij} \Big/ \sum_{j=1}^{N} A_{ij} \tag{5}$$

$$e_k(c) = C_k(c) \Big/ \sum_{c} C_k(c) \tag{6}$$

with:

$$A_{ij} = \sum_{d \in D} 1/(Z^L \cdot P(d \mid M)) \cdot \sum_{t=1}^{L} f_i(t-1) \times a_{ij} \cdot b_j(t) \cdot \sum_{c} e_j(c) \cdot s^t(c) \tag{7}$$

$$C_k(c) = \sum_{d \in} 1/(Z^L \cdot P(d|M)) \cdot \sum_{t=1}^{L} f_i(t) \times b_j(t) \cdot s^t(c) \tag{8}$$

where d are the sequences of the training set D; M is the model; $f_i(t)$ is the probability of emitting the first t vectors of the sequence requiring that the state at the iteration t is k and b_k (t) is the probability of emitting the vectors from the position $t+1$ to the end of the sequence, given that the state at the iteration t is k. The iterative procedure for updating the parameters of the HMM during the training is stopped when the likelihood does not increase (or decrease) any more.

Selecting the topological model

An algorithm based on dynamic programming uses the HMM outputs to locate the transmembrane β-strands along the protein sequence by model optimisation. A similar algorithm was previously used to locate transmembrane α-helices (Jones *et al.*, 1994) and added to the outputs of a neural-network-based method to predict transmembrane β-strands (Jacoboni *et al.*, 2001). The algorithm takes advantage of the notion that transmembrane β-strands in the prototypes of β-barrel membrane proteins are even in number and range from 2 to 22 in the sequence. Briefly, a recursive algorithm generates a scoring matrix for each predicted sequence by evaluating the total sum of the output differences along a segment of fixed length. Minimal and maximal lengths are derived from the database of selected proteins. A model is selected by evaluating the optimal score among those satisfying the observed constraints in the crystals.

For a given sequence position j and for a given model i (i is the number of β-strands) the scoring matrix $\mathbf{S}$ is computed as:

$$\boldsymbol{S}_j^i = \max_{l=\beta_{\min} \to \beta_{\max}} \{s_j^l + \max_{k=j+l+L \to n} \{\boldsymbol{S}_k^{i-1}\}\} \tag{9}$$

Where L and n are the minimum length of a loop segment and the protein length, respectively; s_j^l is the score associated with a transmembrane strand of length l at position j in the sequence.

Topology is then predicted by simply comparing the length of the loops of the two side of the barrel and labelling as extra-cellular the barrel side with the longest loops.

Scoring the prediction

The efficiency of the predictors is scored using the statistical indexes defined in the following.

The HMM accuracy is:

$$Q2 = P/N \tag{10}$$

where P is the total number of correctly predicted residues and N is the total number of residues.

The correlation coefficient C is defined as:

$$C(s) = (p(s) * n(s) - u(s) * o(s))/[(p(s) + u(s))(p(s) + o(s))(n(s) + u(s)) \times (n(s) + o(s))]^{1/2} \tag{11}$$

where, for each class s, p (s) and $n(s)$ are respectively the total number of correct predictions and correctly rejected assignments while $u(s)$ and $o(s)$ are the numbers of under and over predictions.

The accuracy for each discriminated structure s is evaluated as:

$$Q(s) = p(s)/[p(s) + u(s)] \tag{12}$$

where $p(s)$ and $u(s)$ are the same as in Equation 11.

The probability of correct predictions $P(s)$ is computed as:

$$P(s) = p(s)/[p(s) + o(s)] \tag{13}$$

where $p(s)$ and $o(s)$ are the same as in Equation 11.

The segment-based measure (Sov) of the assessment of transmembrane β-strands is computed as previously described (Zemla *et al.*, 1999).

Databases

The training set includes the only 12 non redundant, constitutive β barrel membrane proteins, whose sequences are less than 30% homologous and whose 3D structure have been resolved. The number of β strands forming the transmembrane barrel ranges from 8 to 22. The proteins of the training set are:

- one monomer of the homotrimeric TolC from *Escherichia coli* (1EK9; Koronakis *et al.*, 2000), containing 4 transmembrane β-strands

- protein X from *Escherichia coli* (1QJ8; Vogt and Schulz 1999) and the outer membrane protein A from the same bacterium (1BXW; Pautsch and Schulz, 1998), both with 8 β strands in the barrel;
- phospholipase A from *Escherichia coli* (1QD5; Snijder *et al.*, 1999), with 12 β strands in the barrel;
- porins from *Rhodopseudomonas blastica* (1PRN; Kreusch and Schulz, 1994) and from *Rhodobacter capsulatus* (2POR; Weiss and Schulz, 1992), the Ompf porin from *Escherichia coli* (2OMF; Cowan *et al.*, 1995), the anion selective porin Omp32 from *Comamonas acidovorans* (1E54, Zeth *et al.*, 2000), all with 16 β trands in the barrel;
- maltoporin (2MPR; Meyer *et al.*, 1997) and the sucrose specific porin ScrY (1A0S; Forst *et al.*, 1998), both from *Salmonella typhimurium* and with 18 β strands in the barrel;
- the transporters FhuA (1FCP; Ferguson *et al.*, 1998) and FepA (1FEP; Buchanan *et al.*, 1999) from *Escherichia coli* with 22 β strands in the barrel.

In order to validate the discriminative capability of the predictor three more sets were selected:

- 1239 globular proteins, whose structures have been resolved and whose sequences are less than 25 % similar (http://www.cbrc.jp/papia/papia.html);
- 188 well annotated inner all α membrane proteins (Moller *et al.*, 2000)
- 145 outer membrane proteins from bacteria with sequence identity <30% within the set and with the chains of the training set.

Mapping β barrel membrane proteins on the HMM predictor

The topology of β barrel membrane proteins can be described assigning each residues to one of three types: inner loop, transmembrane β strand, outer loop. The chemico-physical and geometrical characteristics of the three types of segments as deduced by the available structures in the database suggest how to build an HMM for the prediction of the topology of β barrel membrane proteins. The model we develop is shown in Figure 1, and is essentially based on six different types of states, as detailed in the legend for a total of 54 states. The states represented with squares describe the transmembrane β strands while the states shown with circles represent the loops. A statistics on the non-redundant database of outer membrane proteins presently available indicates that the length of the β strands of the training set ranges from 6 to 22 residues (with an average length of 12 residues when considering the length distribution) (Jacoboni *et al.*, 2001). In bacterial outer membrane proteins (so far the only ones present in the data base of available structures) inner loops are generally shorter than outer loops. Furthermore N and C termini of all the proteins lie in the inner side of the membrane. These constraints are modelled by means of the allowed transitions between the states.

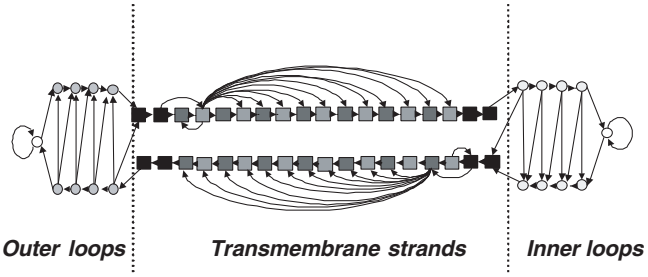

Fig. 1. HMM for the prediction of β-barrel membrane proteins: the states filled with the same colour share the same emission parameters. States represented with squares are labelled as transmembrane states; states represented with circles are labelled as loop states. The connections among the states reflect the allowed transitions.

The different portions of the β barrel membrane proteins have different compositions that reflect the environment, which stabilises the folded protein. In order to model these differences, several sets of emission probabilities have been used. In Figure 1, states depicted with the same colour share the same emission parameters. In particular we used three sets of emission parameters for describing the transmembrane β strands. The first models the edges of β strands, that are rich in large aromatic residues (Schulz, 2000), while the other two attempt to model the presence of residues that alternatively make contacts with the lipid phase and the water phase filling the barrel. The overall number of parameters of this model is 202.

RESULTS AND DISCUSSION

Presently some 720 sequences of outer membrane proteins are annotated in the Swiss Prot database, 626 of which are from bacteria. However in the database of atomic solved structures, if a non-redundant protein set is selected, we are left with 12 structures of constitutive functional outer membrane proteins, all from bacteria. Therefore the HMM has been trained on the constitutive outer membrane proteins that are available, with the algorithms described above. Its architecture (Figure 1) reflects features of the barrel that can derived from the structures of the selected data set (see above). During the training phase, only the paths through the model compatible with the native structure have been taken into account (Krogh, 1994).

The prediction is performed with an '*a posteriori*'

Table 1. Statistical analysis of the predictive performance

	Q2	Q(β)	Q(c)	P(β)	P(c)	C(β)	Sov(β)
HMM Training							
Single sequence	0.83	0.81	0.84	0.80	0.85	0.65	0.80
Multiple sequence	0.84	0.84	0.84	0.81	0.86	0.67	0.85
HMM Testing							
Single sequence	0.76	0.77	0.76	0.72	0.80	0.53	0.64
Multiple sequence	0.83	0.83	0.82	0.79	0.85	0.65	0.83
NN Testing							
Multiple sequence	0.78	0.74	0.82	0.81	0.76	0.56	0.79

$\beta = \beta$ strands, c = non β strands; *NN Testing*: testing is performed with a neural-network-based predictor using as input evolutionary information (Jacoboni *et al.*, 2001). Index definition is given in the System and Methods section.

decoding: the probability for a residue to be in one of the transmembrane states is computed and compared with the probability to be in a loop state (Durbin *et al.*, 1998).

The main novelty of the model is the implementation of the input in the form of sequence profiles as detailed above. To our knowledge this is new for HMM based models used to predict the topology of membrane proteins. Sequence profiles have been computed from the alignments as derived with PSI-BLAST (Altschul *et al.*, 1997) on the non-redundant database (ftp://ftp.ncbi.nlm.nih.gov/blast/db/). The results scored with statistical indexes are shown in Table 1. The performance is tested with a jack-knife validation procedure. For sake of comparison the results obtained for the same predictive task with an HMM based on single sequence are also listed. It appears that the inclusion of sequence profile greatly improves the predictive capability of the model. Furthermore results obtained on the same set with a neural network method recently implemented (Jacoboni *et al.*, 2001) are also shown. It is evident that when the per residue performance is evaluated, the HMM model trained on the sequence profile is performing better than the neural network predictor having the same input code.

A typical output of the predictor is shown in Figure 2. The probability signal associated with each residue along the protein sequence is however somewhat blurred and not sufficient to give a correct topology prediction for all the proteins in the data set. To regularise the results a model optimisation method (based on a dynamic programming algorithm) is used. By this, the number of proteins whose topology was correctly predicted by the neural network based method increased from 6 to 8 out of 11 proteins included in the training set (Jacoboni *et al.*, 2001). Similarly the number of protein topologies correctly predicted by the HMM model increases from 5 to 7, out of the 12 proteins of the training set (data not shown).

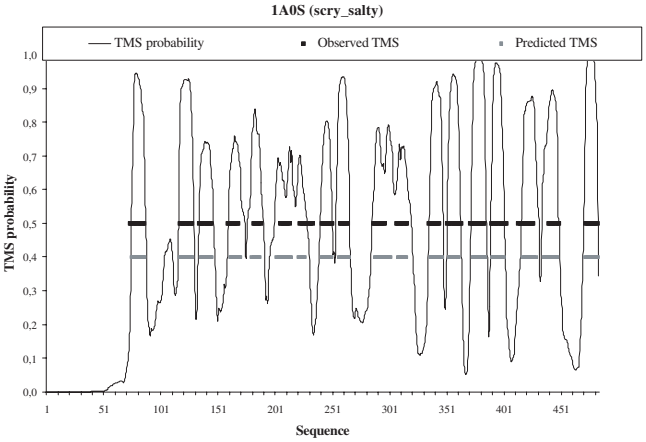

Fig. 2. Prediction of the transmembrane strands of the sucrose specific porin ScrY (1A0S; Forst *et al.*, 1998), from *Salmonella typhimurium* and with 18 β strands in the barrel. The '*a posteriori*' probability for transmembrane states is plotted along the sequence. Grey segments represent the transmembrane β strands of the model optimised with the dynamic programming algorithm. Black segments represent the β strands observed in the structure resolved at atomic resolution.

Discriminative power

We tested the capability of the HMM model to discriminate outer membrane proteins from other protein types. To this purpose we filtered with the model different sets of protein chains, comprising annotated outer membrane proteins, all helical membrane proteins and globular proteins. For each sequence the probability $P(s \mid M)$ of being emitted by the model is computed. This value depends on the length L of the sequence s. An index $I(s \mid M)$ is computed in order to normalise the L dependence:

$$I(s \mid M) = -1/L \log P(s \mid M) \qquad (14)$$

For each set of predicted proteins a distribution of the correctly accepted and rejected proteins is made as a function of the $I(s \mid M)$ value. By averaging over the $I(s \mid M)$ maximum values, a hypothetical threshold value of 2.86 is obtained. In Figure 3 the percentage of the chains classified as outer membrane proteins is plotted as a function of the $I(s \mid M)$ value for the all the protein sets. At this threshold value, 84% of the well annotated outer membrane proteins of the Swiss Prot data with sequence identity <30% to those of the training set are accepted. Concomitantly the rate of false positive is 10% both for globular proteins and all helical membrane proteins. Below the threshold value, the rate of false positives decreases. However, the percentage of outer membrane proteins not accepted increases, indicating that the model is accurate enough to capture the basic features of most

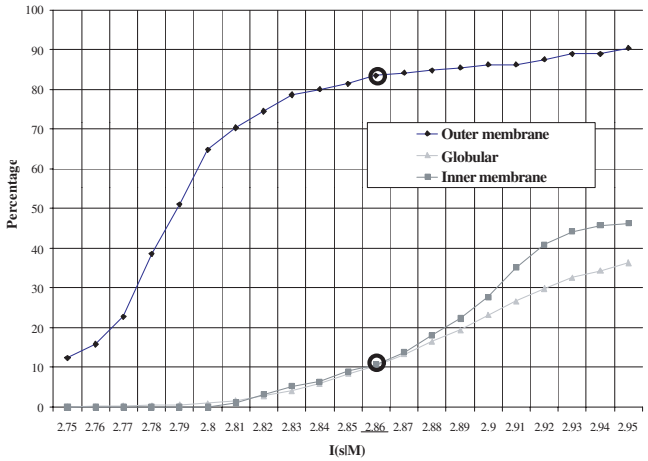

Fig. 3. Discriminative capability of the HMM. The cumulative distribution of the percentage of the proteins correctly classified outer membrane proteins by the model is plotted as a function of the index $I(s \mid M) = -1/L \log P(s \mid M)$ (see text for details). At increasing value of the index the percentage of outer membrane proteins correctly classified increases as well as the percentage of false positives (globular and all helical inner membrane proteins). When the threshold value is set at 2.86, 84% of the protein are correctly accepted as outer membrane proteins, with a rate of false positives of 10%.

but not all of the outer membrane proteins. This is possibly due to the paucity of structures presently available.

CONCLUSIONS

In this paper we exploit the capability of an HMM predictor to model constitutive outer membrane proteins known at atomic resolution. We develop an HMM predictor which derives information from sequence profile after multiple sequence alignment. Furthermore we analyse the applicability of the tool to large-scale protein sequence analysis to select outer membrane proteins.

In spite of the paucity of the structures presently available, the model captures most of the distinguished features of outer membrane proteins and in this it compares favourably with one other method based on neural networks recently developed to solve the same task (Jacoboni *et al.*, 2001). Additionally the HMM model can discriminate between soluble, all helical membrane and outer membrane proteins, suggesting its application to complete genomes.

AKNOWLEDGEMENTS

This work was partially supported by a grant of the Ministero della Università e della Ricerca Scientifica e Tecnologica (MURST) for the project 'Hydrolases from Termophiles: Structure, Function and Homologous and Heterologous Expression' and by a grant for a target project in Biotechnology of the Italian Centro Nazionale delle Ricerche (CNR), both delivered to R.C. R.C. acknowledges also the EC grant Biowulf IST 1999-20232 for supporting the development of DNCBLAST, a parallelized version of PSI-BLAST for P.C. nets. A.K. is supported by a grant from the National Danish Research Foundation.

REFERENCES

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acid Res.*, **25**, 3389–3402.

Buchanan,S.K., Smith,B.S., Venkatramani,L., Xia,D., Esser,L., Palnitkar,M., Chakraborty,R., van der Helm,D. and Deisenhofer,J. (1999) Crystal structure of the outer membrane active transporter FepA from *Escherichia coli*. *Nat. Struct. Biol.*, **6**, 56–63.

Cowan,S.W., Garavito,R.M., Jansonius,J.N., Jenkins,J.A., Karlsson,R., Konig,N., Pai,E.F., Pauptit,R.A., Rizkallah,P.J., Rosenbusch,J.P., Rummel,G. and Schirmer,T. (1995) The structure of OmpF porin in a tetragonal crystal form. *Structure*, **3**, 1041–1050.

Durbin,R., Eddy,S., Krogh,A. and Mitchinson,G. (1998) *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge Univ. Press, Cambridge.

Ferguson,A.D., Hofmann,E., Coulton,J.W., Diederichs,K. and Welte,W. (1998) Siderophore-mediated iron transport: crystal structure of FhuA with bound lipopolysaccharide. *Science*, **282**, 2215–2220.

Forst,D., Welte,W., Wacker,T. and Diederichs,K. (1998) Structure of the sucrose-specific porin ScrY from *Salmonella typhimurium* and its complex with sucrose. *Nat. Struct. Biol.*, **5**, 37–46.

Jacoboni,I., Martelli,P.L., Fariselli,P., De Pinto,V. and Casadio,R. (2001) Prediction of the transmembrane regions of beta-barrel membrane proteins with a neural network-based predictor. *Protein Sci.*, **10**, 779–787.

Jones,D.T., Taylor,W.R. and Thornton,J.M. (1994) A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry*, **33**, 3038–3049.

Koronakis,V., Sharff,A., Koronakis,E., Luisi,B. and Hughes,C. (2000) Crystal structure of the bacterial membrane protein TolC central to multidrug efflux and protein export. *Nature*, **405**, 914–919.

Kreusch,A. and Schulz,G.E. (1994) Refined structure of the porin from *Rhodopseudomonas blastica.* Comparison with the porin from *Rhodobacter capsulatus*. *J. Mol. Biol.*, **243**, 891–905.

Krogh,A. (1994) Hidden Markov models for labeled sequences. In *Proceedings 12th International Conference on Pattern Recognition*. IEEE Computer Society Press, Singapore, pp. 140–144.

Krogh,A., Larsson,B., von Heijne,G. and Sonnhammer,E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.

Meyer,J.E., Hofnung,M. and Schulz,G.E. (1997) Structure of maltoporin from *Salmonella typhimurium* ligated with a nitrophenyl-maltotrioside. *J. Mol. Biol.*, **266**, 761–775.

Moller,S., Kriventseva,E.V. and Apweiler,R. (2000) A collection of well characterised integral membrane proteins. *Bioinformatics*, **16**, 1159–1160.

Pautsch,A. and Schulz,G.E. (1998) Structure of the outer membrane protein A transmembrane domain. *Nat. Struct. Biol.*, **5**, 1013–1017.

Rost,B., Casadio,R., Fariselli,P. and Sander,C. (1995) Transmembrane helices predicted at 95% accuracy. *Protein Sci.*, **4**, 521–533.

Rost,B., Fariselli,P. and Casadio,R. (1996) Topology prediction for helical transmembrane proteins at 86% accuracy. *Protein Sci.*, **5**, 1704–1718.

Schulz,G.E. (2000) β-barrel membrane proteins. *Curr. Opin. Struct. Biol.*, **10**, 443–447.

Snijder,H.J., Ubarretxena-Belandia,I., Blaauw,M.I., Kalk,K.H., Verheij,H.M., Egmond,M.R., Dekker,N. and Dijkstra,B.W. (1999) Structural evidence for dimerization-regulated activation of an integral membrane phospholipase. *Nature*, **401**, 717–721.

Tusnady,G.E. and Simon,I. (1998) Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *J. Mol. Biol.*, **283**, 489–506.

Vogt,J. and Schulz,G.E. (1999) The structure of the outer membrane protein Ompx from *Escherichia Coli* reveals mechanisms of virulence. *Structure*, **7**, 1301–1309.

Weiss,M.S. and Schulz,G.E. (1992) Structure of porin refined at 1.8 Å resolution. *J. Mol. Biol.*, **227**, 493–509.

Zemla,A., Venclovas,C., Fidelis,K. and Rost,B. (1999) A modified definition of Sov, a segment-based measure of protein secondary structure prediction assessment. *Proteins*, **34**, 220–223.

Zeth,K., Diederichs,K., Welte,W. and Engelhardt,H. (2000) Crystal structure of Omp32, the anion-selective porin from *Comamonas acidovorans*, in complex with the a periplasmic peptide at 2.1 Å resolution. *Structure*, **8**, 981–992.

APPENDIX

The aim of this section is to demonstrate that:

$$Z = (S^A/A!) \cdot \sum_c e_k(c) \tag{15}$$

This result proves that the normalisation constant Z is independent of the values of the vector components. It is possible to consider it as a constant and to implement the dynamic programming algorithm and the training procedure as described in the text.

The following expression makes explicit the constraints of the integral expression for Z of Equation 3. The index t is omitted for sake of simplicity.

$$\begin{aligned} Z = \int_0^S \mathrm{d}s(1) \int_0^{S-s(1)} \mathrm{d}s(2) \ldots \int_0^{S-s(1)-s(2)-\ldots s(A-2)} \\ \times \mathrm{d}s(A-1) \\ \times [s(1) \cdot e_k(1) + s(2) \cdot e_k(2) \\ + \cdots + s(A-1) \cdot e_k(A-1) \\ + (S - s(1) - s(2) - \cdots - s(A-1)) \cdot e_k(A)] \end{aligned} \tag{16}$$

Expression of Equation 16 is a multiple integral of $(A-1)$ independent variables: the variable $s(A)$ is cancelled by means of the normalisation condition.

Two lemmas will be demonstrated in advance in order to simplify the final proof.

LEMMA 1. *We want to demonstrate the following expression:*

$$\begin{aligned} I_A \equiv \int_0^S \mathrm{d}s(1) \int_0^{S-s(1)} \mathrm{d}s(2) \ldots \int_0^{S-s(1)-s(2)-\ldots s(A-1)} \\ \times \mathrm{d}s(A) = S^A/A! \end{aligned} \tag{17}$$

PROOF. The proof proceeds by complete induction on the number of variables, A.

- $$I_1 = \int_0^S \mathrm{d}s(1) = S \tag{18}$$
- If the expression of Equation 18 is true for A, then for $(A+1)$ the value of integral is:

$$\begin{aligned} I_{A+1} = \int_0^S \mathrm{d}s(1) \int_0^{S-s(1)} \mathrm{d}s(2) \ldots \int_0^{S-s(1)-s(2)-\ldots s(A)} \\ \times \mathrm{d}s(A+1) \end{aligned} \tag{19}$$

  Using the inductive hypothesis to compute the last A integrals and substituting the constant S with $(S-s(1))$ we obtain:

$$I_{A+1} = (1/A!) \int_0^S (S-s(1))^A \mathrm{d}s(1) = S^{A+1}/(A+1)! \tag{20}$$

This result is the proof of the lemma.

LEMMA 2. *Using the result of Lemma 1 it can be demonstrated that:*

$$\begin{aligned} \int_0^S s(1)\mathrm{d}s(1) \int_0^{S-s(1)} \mathrm{d}s(2) \ldots \int_0^{S-s(1)-s(2)-\ldots s(A-1)} \\ \times \mathrm{d}s(A) = S^{A+1}/(A+1)! \end{aligned} \tag{21}$$

PROOF. Using the following variable exchange:

$$y = S - s(1) \tag{22}$$

the first member of Equation 21 become:

$$\begin{aligned} \int_0^S (S-y)\mathrm{d}y \int_0^y \mathrm{d}s(2) \ldots \int_0^{y-s(2)-\ldots s(A-1)} \mathrm{d}s(A) \\ = \int_0^S (S-y) \cdot (y^{A-1})/(A-1)!\mathrm{d}y \\ = (S^{A+1})/(A+1)! \end{aligned} \tag{23}$$

Lemma 1 is applied to the last $(A-1)$ variables, upon substitution of S with y. The result of Equation 23 is the proof of the lemma.

Proof of Equation 15. The proof proceeds by complete induction on the number A of variables, starting from 2.

$$A = 2 \Rightarrow Z = \int_0^S \mathrm{d}s(1)[s(1) \cdot e_k(1) + (S - s(1)) \cdot e_k(2)] = S^2 \cdot (e_k(1) + e_k(2))/2 \quad (24)$$

- If expression 15 is correct for A, then for $(A + 1)$ the integral become:

$$\begin{aligned} Z = &\int_0^S \mathrm{d}s(1) \int_0^{S-s(1)} \mathrm{d}s(2) \ldots \int_0^{S-s(1)-s(2)-\ldots s(A-1)} \\ &\times \mathrm{d}s(A)[s(1) \cdot e_k(1) + s(2) \cdot e_k(2) \\ &+ \cdots + s(A) \cdot e_k(A) \\ &+ (S - s(1) - s(2) - \cdots - s(A)) \cdot e_k(A + 1)] \\ = &\int_0^S s(1) \cdot e_k(1) \mathrm{d}s(1) \int_0^{S-s(1)} \mathrm{d}s(2) \ldots \\ &\times \int_0^{S-s(1)-s(2)-\ldots s(A-1)} \mathrm{d}s(A) \\ &+ \int_0^S \mathrm{d}y \int_0^y \mathrm{d}s(2) \ldots \int_0^{y-s(2)-\ldots s(A-1)} \mathrm{d}s(A) \\ &\times [s(2) \cdot e_k(2) + \cdots + s(A) \cdot e_k(A) + (y - s(2) \\ &- \cdots - s(A)) \cdot e_k(A + 1)] \end{aligned} \quad (25)$$

The first term of the second member of Equation 25 is solved by Lemma 2. Concerning the second term, we can use the variable exchange of Equation 22 and then we can apply the inductive hypothesis. It results that:

$$Z = S^{A+1}/(A + 1)! \cdot \sum_{c=1}^{A+1} e_k(c) \quad (26)$$

Equation 26 proves Equation 15.