First publ. in: Protein Science 7 (1998), pp. 2413-2420

Prediction by a neural network of outer membrane β-strand protein topology

KAY DIEDERICHS, JÖRG FREIGANG, STEPHAN UMHAU, KORNELIUS ZETH, AND JASON BREED

Universität Konstanz, Fakultät für Biologie (M656), D-78457 Konstanz, Germany

Abstract

An artificial neural network (NN) was trained to predict the topology of bacterial outer membrane (OM) β-strand proteins. Specifically, the NN predicts the *z*-coordinate of Cα atoms in a coordinate frame with the outer membrane in the *xy*-plane, such that low *z*-values indicate periplasmic turns, medium *z*-values indicate transmembrane β-strands, and high *z*-values indicate extracellular loops. To obtain a training set, seven OM proteins (porins) with structures known to high resolution were aligned with their pores along the *z*-axis. The relationship between Cα *z*-values and topology was thereby established. To predict the topology of other OM proteins, all seven porins were used for the training set. *Z*-values (topologies) were predicted for two porins with hitherto unknown structure and for OM proteins not belonging to the porin family, all with insignificant sequence homology to the training set. The results of topology prediction compare favorably with experimental topology data.

Keywords: beta-strand prediction; cross-validation; neural network; outer membrane protein; porin; topology prediction

Outer membrane (OM) proteins are functional and structural constituents of the OM of Gram-negative bacteria. Transport across the OM is an essential first stage of nutrient uptake into bacteria. OM proteins participate in transport complexes with phosphotransferase or ATP-binding cassette transporter systems. Structurally known integral membrane proteins not from the OM, such as the photosynthetic reaction center (Deisenhofer et al., 1985), cytochrome *c* oxidase (Ostermeier et al., 1996; Yoshikawa, 1997), or rhodopsin (Schertler et al., 1993; Unger et al., 1997) are mainly composed of transmembrane (TM) α-helices. However, there is growing evidence that many OM proteins fall into a different folding class, being mainly composed of TM β-strands (Cowan & Rosenbusch, 1994). Porins are OM proteins that function as water-filled channels. The atomic structures of several porins are known, but structural data are not yet available for a large number of nonporin OM proteins. The database of genetic and functional data for such proteins is considerable and structural models may help to interpret these data and to effectively plan further experimentation. They could be used in conjunction with experimental low-resolution topology mapping procedures, such as monoclonal antibody binding and epitope mapping or by genetically creating fusion proteins, e.g., for insertion of protease cleavage sites (Ehrmann et al., 1997). Methods for predicting the topology of OM proteins based on their sequences are thus required.

Previous topology predictions concentrated on predicting the locations of β-strands within the amino acid sequence. In contrast to TM α-helices, which may be predicted to high accuracy (Rost et al., 1995), this task appears to be more difficult for TM β-strands. Methods for the prediction of TM β-strands in OM proteins based on physicochemical properties of their amino acid composition were proposed by Paul and Rosenbusch (1985), Vogel and Jähnig (1986), Jähnig (1990), Welte et al. (1991), and Schirmer and Cowan (1993). Although accurate results can be obtained if used in conjunction with a variety of sources of experimental information (Tommassen, 1988), these methods generally meet with limited success as TM β-strands in OM proteins are amphipathic and thus more difficult to find than blocks of hydrophobic residues. A simple alternating pattern of hydrophobic and hydrophilic residues β-strands is also insufficient for TM β-strand identification as the amphipathicity may be more complex. Pore-lining residues are not necessarily hydrophilic. Hydrophobic pore-lining residues may interact with hydrophobic residues in loops (e.g., loop 3 in nonspecific porins) or with transported solutes, as in maltoporin. A rule-based approach for identifying TM β-strands in porins was proposed by Gromiha et al. (1997) and was shown to be effective within a limited set of three nonspecific porin structures. Rules giving the directions of the strands were not established, neither was the same

Reprint requests to: Kay Diederichs, Universität Konstanz, Fakultät für Biologie (M656), D-78457 Konstanz, Germany; e-mail: Kay.Diederichs@uni-konstanz.

rule set applied to specific, 18-stranded porins or to the superfamily of OM proteins.

An artificial neural network (NN) may be used to identify and model complex patterns in biological data (Presnell & Cohen, 1993). NNs have been successfully employed for the analysis or prediction of a number of protein properties, including secondary structure (Qian & Sejnowski, 1988; Holley & Karplus, 1989; Kneller et al., 1990; Muskal & Kim, 1992; Geourjon & Deléage, 1994; Rost & Sander, 1994a), solvent accessibility (Rost & Sander, 1994b), folding class (Dubchak et al., 1993), ATP-binding motifs (Hirst & Sternberg, 1991) and side-chain packing (Milik et al., 1995). Here, we apply an NN to the task of predicting the z-values of Cα atoms of OM proteins. This prediction addresses both location and direction of β-strands.

A feedforward NN consists of two or more layers of processing units. The first and last layers are termed input layer and output layer, respectively, and intermediate ones are called hidden layers (Fig. 1A). Each unit i of a given layer is connected to each unit j of the next layer; the strength of each connection is given by a weight w_{ij}. The state s of each unit is a value in the range 0–1. Input units are assigned their states directly from the given input data, whereas the states s_j of higher layers j are computed by

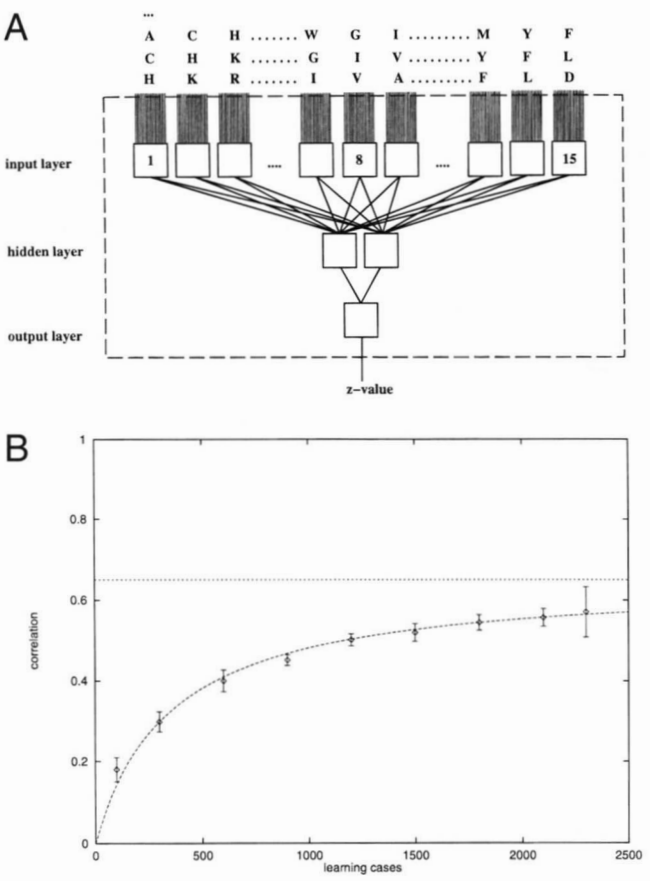

Fig. 1. A: Architecture of NN (schematical). Examples for input amino acid sequences (one-letter code) resulting from the NN input window sliding along the protein sequence are given (top). **B:** Cross-validated correlation of predicted and observed z-values as a function of the number of training cases. Error bars indicate standard deviations of correlation values obtained by different assignments of available cases into training and test cases. The dashed curve represents a nonlinear fit (Equation 4) to the data.

$$s_j = \frac{1}{1 + e^{-\sum_i w_{ij} s_i + w_{j0}}} \tag{1}$$

where w_{j0} is a bias from the states s_i of lower layers.

During the "training phase" of NN operation, a "training set" of cases, each describing the states s_i of the input units and their associated output values o_k, is presented to the NN. The s_i and o_k of these cases are usually experimentally known quantities and are normalized to lie within the 0–1 range. In a feedback procedure (Rumelhart et al., 1986), all weights $\{w\}$ are then iteratively adjusted such that the total error, given by the sum of squared differences between the computed states s_k of output unit and these target output values o_k

$$E\{w\} = \sum_{cases} \sum_k (s_h - o_k)^2 \tag{2}$$

is minimized.

Once the NN has been adjusted (or "trained") during the "training phase," the weights reflect some of the properties of the underlying problem. In the "prediction phase," new cases with known input values, but unknown output values, are fed into the NN and the previously obtained weights $\{w\}$ are used to calculate the output states s_k, which are then taken as predictions for the true o_k.

Results

Cross-validation of NN performance

Figure 1B gives the dependence of prediction success for the test cases as a function of the number of training cases available in the training phase. As a measure of prediction success, we chose the correlation coefficient c

$$c = \frac{\sum x_i y_i - \frac{1}{n} \sum x_i \sum y_i}{\left(\sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2\right)^{1/2} \left(\sum y_i^2 - \frac{1}{n}\left(\sum y_i\right)^2\right)^{1/2}} \tag{3}$$

between predicted and observed z-values. C should be close to 0.0 if no training cases are present, and ≤ 1.0 if the number of training cases n goes to infinity. A simple saturation-type model for the value of the cross-validated correlation c as a function of the number of available training cases n meeting these two requirements is given by

$$c = \frac{n}{\frac{n}{a} + b} \tag{4}$$

with adjustable parameters a, $0 \leq a \leq 1$ as the saturation value (for $n \to \infty$) of the correlation and b, $b > 0$ as an offset.

Indeed, this function yielded a satisfactory fit (dashed line) to the results obtained for training sets of different size, and values of ($a = 0.65$, $b = 541$) were obtained. Such results signify that with the current architecture of the NN a maximum correlation around 0.65 can be obtained for $n \to \infty$. However, even with the seven porin structures presently available ($n = 2{,}388$), the correlation is 0.58.

When calculations were repeated after deleting PhoE from the training set, correlation coefficients were within the error bars shown in Figure 1B, indicating that the influence of homology on the cross-validation results obtained with all seven porins is minor.

Predictions of known structures: Comparison of predicted and actual topologies

To demonstrate the agreement between prediction and experimental structure, we present the actual and predicted topologies of porin from *Rhodobacter capsulatus* in a "topology plot" (Fig. 2). The correlation coefficient c in this case was determined by cross-validation to be 0.50. The highest pairwise identity of *R. capsulatus* porin to any member of the training set is 23% (*Paracoccus* porin); average pairwise identity is 14.8%. Following are a number of salient features: A predicted z-value of 0.6 or less generally corresponds to a TM or periplasmic location. The signal for periplasmic turns is somewhat weak, with predicted values between 0.2 and 0.4. In contrast, the signals for extracellular loops are strong, either matching or exceeding the actual z-values in most cases. The prediction differs significantly from the actual topology in two regions. The first of these is residues 85–115, corresponding to the pore-constricting loop 3, which folds back into the β-barrel. The NN predicts an extracellular loop for this region; thus, it identifies an underlying pattern that is masked by highly specific interactions between loop 3 and the barrel wall, which are not extracted from the sequence. The second area of disagreement covers residues 210–230. In porin from *R. capsulatus*, this region forms a small globular extracellular domain that is unique among porins of known structure. The NN prediction is for two TM β-strands and a tight periplasmic turn in this area. The presence of hydrophobic residues in the core of the globular extracellular domain may give this region a more amphipathic character than is usual for extracellular loops in OM proteins, and thus lead to the prediction of a TM localization for these residues. However, the overall correspondence between prediction and experimental data is good and confirms that the general topological pattern of OM proteins can be predicted by this method, with the qualification that unique and specific patterns may be overlooked.

Negative controls

If the network is to be able to correctly predict the topology of OM β-barrel proteins, it must also be able to distinguish between membrane proteins with α-helical or with β-sheet secondary structure. As a control, we used the sequence of human rhodopsin, a membrane protein known to be composed of α-helices (Schertler et al., 1993; Unger et al., 1997). We had expected the predicted z-values randomly fluctuating around 0.5. The actual output is markedly different both from our expectations and from the outputs for porin: sharp transitions from low to high to low predicted z-values alternate with broad stretches of low predicted z-values. Similar outputs were obtained for other α-helix-containing membrane proteins of known structure, i.e., cytochrome c oxidase (Ostermeier et al., 1996; Yoshikawa, 1997) and photosynthetic reaction center (Deisenhofer et al., 1985). Thus, NN output for TM α-helix proteins is readily differentiated from output for TM β-strand proteins.

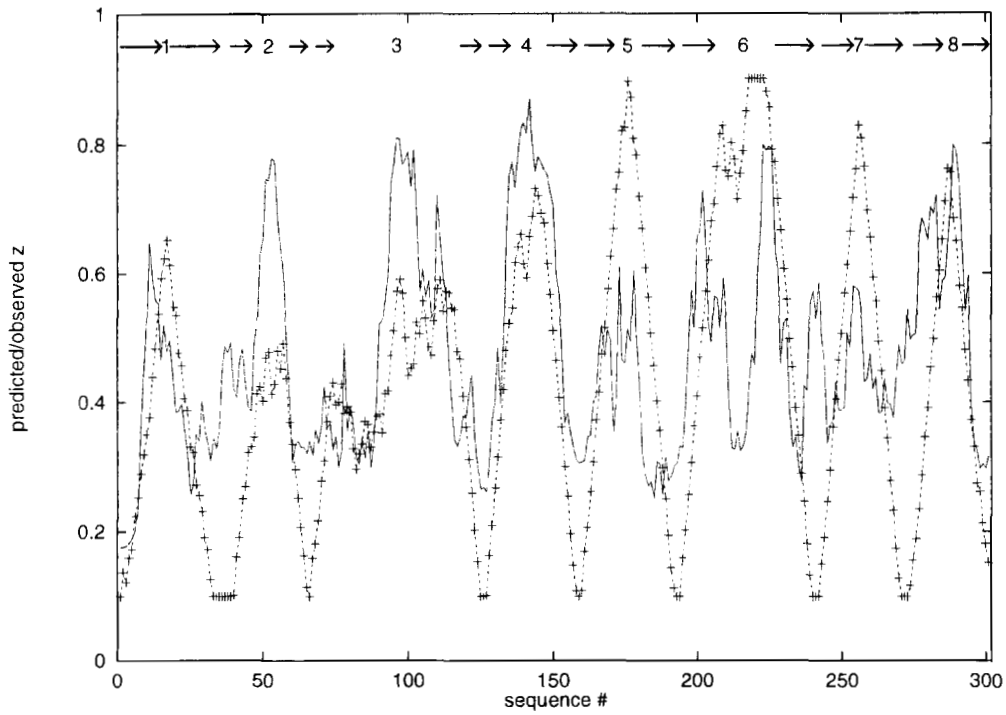

Fig. 2. Topology plot: predicted z-values are plotted against the amino acid sequence. This example shows a comparison of actual (dashed line with markers) and predicted (continuous line) topologies for porin from *R. capsulatus*. Predicted z-values were obtained without information from the *R. capsulatus* structure in the training set. Arrows correspond to the 16 β-strands revealed in the X-ray structure. N- and C-termini are periplasmic.

Further negative controls were performed with a number of soluble proteins. None of the controls exhibited periodic alterations between high and low z values, as is typical for porins. In all cases, topology predictions resembled plots of a random variable with mean 0.5.

Topology predictions of proteins with unknown structure

The set of proteins we investigated was chosen on the basis of substantial available experimental topological data (OmpA, Hib porin, FhuA) or of the prospect of structure determination in the near future (OmpA, Omp32, FepA, FhuA). These proteins also represent systems for which topology prediction may be of interest to microbiologists or molecular biologists in the field.

OmpA from Escherichia coli

OmpA is one of the major protein constituents of the outer membrane of *Escherichia coli*. It participates in the maintenance of cell surface integrity (Sonntag et al., 1978) and is a receptor for various phages and colicins (Morona et al., 1985). Although there is no crystal structure for OmpA yet, a number of its general structural features have been determined and a structural model proposed (Morona et al., 1984; Vogel & Jähnig, 1986; Ried et al., 1994; Koebnik & Krämer, 1995; Koebnik, 1996), with an N-terminal domain (residues 1–170) forming an eight-stranded TM β-barrel and the remaining residues forming a C-terminal periplasmic domain. An alternative model of a 16-stranded β-barrel for OmpA has more recently been put forward (Stathopoulos, 1996).

Before comparing our prediction to the competing models for OmpA, we first used experimental data on the localization of specific residues to assess the reliability of the NN output (Chen et al., 1980; Morona et al., 1984, 1985; Freudl et al., 1986, 1989; Ried et al., 1994; Ruppert et al., 1994; Georgiou et al., 1996). Experimental and predicted localizations are in agreement in the majority of cases (13 out of 19; see Table 1 in the supplementary material). Thus, it seems that the predicted topology for OmpA agrees mostly with the experimental data. The output of the NN and its interpretation as a topological model are given in Figure 3A. Four "extracellular" peaks are evident in the N-terminal domain, suggesting eight TM β-strands between residues 1 and 170. However, the strongest "extracellular" peak occurs after residue 170 and is followed by a series of peaks that, overall, result in a prediction of at least 16 TM β-strands for OmpA. Thus our prediction tends to support the more recent model of OmpA from Stathopoulos (1996).

Porin from Haemophilus influenzae Type b

Haemophilus influenzae (Hib) is a Gram-negative bacterium and a cause of infant meningitis. The nonspecific porin (341 residues, 38 kDa) from Hib is well characterized (Srikumar et al., 1992a, 1992b, 1997). Its active form is trimeric and it forms mildly anion selective channels with mean conductance (1.1 nS) and mass exclusion limit (1,400 Da) values significantly higher than those of nonspecific porins of known structure. Its topology has been investigated using monoclonal antibodies to four epitopes between 6 and 11 residues long. Antibody binding (in whole cells, and to porin in micelles) was determined using flow cytometry and ELISA techniques (Srikumar et al., 1992a, 1992b). Two epitopes were strongly bound: residues 162–172 and 318–325. The first of these regions gives a strong signal for an extracellular location in the topology prediction for Hib porin (Fig. 3B), whereas the second is predicted to have a TM location. A third epitope, residues 148–153, was not available for antibody binding. A periplasmic location is predicted for this sequence. A fourth epitope, residues 112–126, was also not bound by antibodies. This epitope corresponds to the proposed gating loop for Hib porin, which in all known porin structures is the third extracellular loop. Our prediction places this sequence within loop 3. As in all porins of known structure, loop 3 of Hib porin is likely to fold back into the pore and is thus not available at the extracellular surface for antibody binding. The C3 epitope of the poliovirus VP1 capsid protein was introduced into Hib porin after residue 174; flow cytometry using anti-C3 antibodies indicated that the epitope had an extracellular location (Srikumar et al., 1997). Residues 174 and 175 are predicted to reside in the fourth extracellular loop. Overall, there is good agreement between predicted and experimental locations for specific residues in Hib porin.

Omp32

Omp32 is the major protein component of the OM of *Comamonas acidovorans* and is an anion-selective nonspecific porin (332 residues, 36 kDa). This porin has been crystallized (Zeth et al., 1998) and a structure determination should be complete within the near future, which will provide an excellent assessment of topology prediction accuracy of the NN. A topological model of Omp32 based on the NN predictions is given in Figure 3C. The topology of Omp32 was analyzed by proteolysis (Gerbl-Rieger et al., 1992). Six locations were found to be protease accessible. Two of these six locations (residues 123–126 and 200–201) are predicted to be extracellular, with a further two (26–30 and 317) predicted to be at the TM/extracellular interface. Thus, in four cases out of six, prediction and experimental data may be reconciled. The forthcoming structure of Omp32 will give a more thorough test of prediction performance for this porin.

FepA and FhuA

Both FepA (724 residues, 78 kDa) and FhuA (714 residues, 79 kDa) are involved in the TonB-dependent, active uptake of iron-bearing siderophores into *E. coli*. They both reside in the OM and bind ferric enterobactin and ferrichrome, respectively, allowing passive uptake of these siderophores into the periplasmic space; from there they are actively transported across the inner membrane. Both proteins are predicted to form antiparallel β-barrels of up to 32 strands (Murphy et al., 1990; Koebnik & Braun, 1993). Much topological information is available for both proteins. In addition, both proteins have been crystallized (Buchanan et al., 1996; Ferguson et al., 1998) and the future comparison of structure and prediction for these nonporin proteins will greatly aid the evaluation of the NN.

The topology of FepA was investigated using monoclonal antibodies to epitopes between 10 and 50 residues in length (Murphy et al., 1990). Five epitopes: residues 27–37, 204–227, 258–290, 290–339, and 382–400 were reported to show antibody binding by flow cytometry and, thus, assumed to have an extracellular location. Due to the length of these sequence stretches, it is difficult to correlate the experimental data with the prediction. A stretch of 40 residues may well have both extracellular and periplasmic as well as TM components. However, all epitopes contain sections with strong extracellular predictions and thus may have extracellular regions sufficient to account for antibody binding (Murphy et al., 1990). Results of a site-directed spin-labeling study of FepA suggested that residue 280 is surface-localized and that residue 310 is

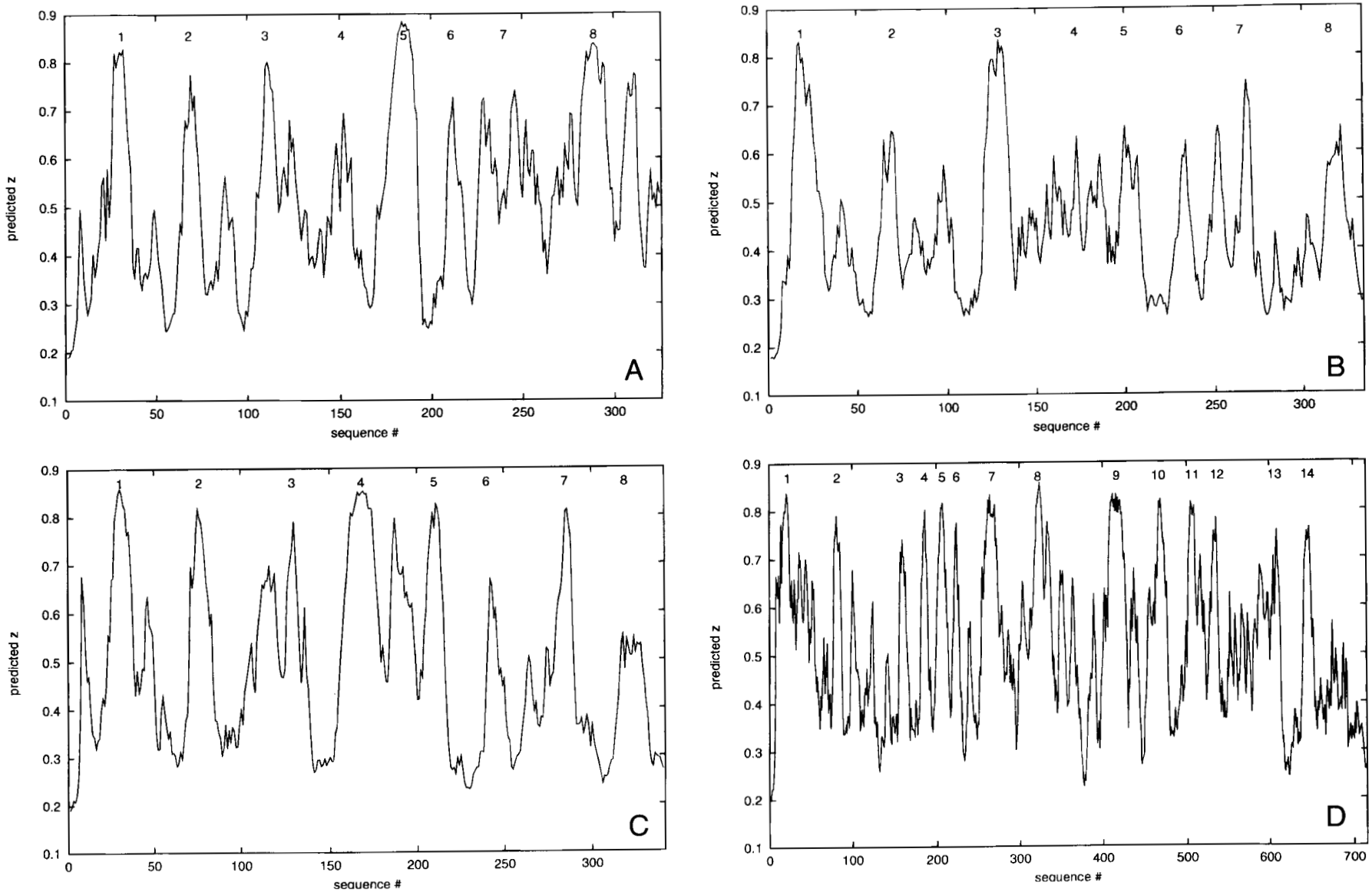

Fig. 3. Topology plot for (**A**) OmpA from *E. coli*, (**B**) porin from *H. influenzae* Type b, (**C**) porin from *C. acidovorans* (Omp32), (**D**) FhuA from *E. coli*. Numbers above the plots denote positions of NN-predicted extracellular loops with high *z*-values.

located close to the extracellular-membrane interface (Liu et al., 1994). In our prediction, these residues have an extracellular and an extracellular/TM localization, respectively. A more recent spin-labeling study assigns a TM localization to residues 245, 249, and 253 (Klug et al., 1997). All three residues have a TM localization in our prediction. A double mutagenesis study showed that residues 286 and 316 are involved in siderophore binding and the action of colicins B and D; they should thus be extracellular (Newton et al., 1997). We predict that these residues are TM and extracellular, respectively. In general, the agreement between predicted and experimental localizations is good for FepA.

FhuA has been subject to a number of topological investigations. Peptide insertion mutants were generated and their expression and functionality observed (Koebnik & Braun, 1993). Those mutants that retained function were deemed not to have been altered in a TM β-strand. Active mutants were, therefore, altered in a loop or turn region. Active mutants with reduced phage or colicin sensitivity are assumed to have an extracellular location (Koebnik & Braun, 1993). Experimental and predicted localizations were identical in 15 of 17 cases. Residues 301, 405, and 417 were determined to be extracellular by the use of polio C3 reporter epitope and flow cytometry (Moeck et al., 1994). All three residues are predicted to be extracellular. A topological analysis of FhuA using monoclonal antibodies indicated extracellular or periplasmic localizations for a number of residue stretches (Moeck et al., 1995). Again, the length of the peptide stretches used as epitopes in that study makes interpretation more difficult as 35 or more residues may well encompass more than one topological location. However, each epitope is predicted to have, at least partly, the same localization as found using monoclonal antibodies, e.g., determinant sequences 321–381 and 417–550 were bound by monoclonal antibodies with an extracellular localization and both sequences contain stretches predicted to be extracellular (348–371, 434–461, 487–510, and 534–550, respectively). The comparison of FhuA experimental and predicted topology is, overall, most encouraging (see Table 2 in the supplementary material).

Discussion

An impression of the quality of the NN topology prediction can most easily be obtained from Figure 2. It is obvious that although the prediction of z-coordinates is far from perfect, the essential features of the porin fold are captured. It should be noted that no filtering (e.g., constraining differences betwen Cα z-values to a maximum of 3.8 Å) or smoothing was applied to the output. In our experience, qualitative aspects of topology plots, such as direction and locations of β-strand residues, are best derived by visual inspection. It appears that the position of the residues with highest z-value and the directions of the β-strands are predicted quite reliably, whereas the number of residues in a TM strand is not easily inferred from the topology plot.

Efficient abstraction of rules by an NN requires that the number of training cases be substantially higher than the number of independent variables, and, for a given architecture of the NN, its accuracy generally increases with the number of data available (Chandonia & Karplus, 1996). To overcome the paucity of available structures of OM proteins, we characterized the asymptotic learning behavior of a simple NN architecture by cross-validation. These results are valid because the sequence homology between training and test data is generally low, and were confirmed by exclusion of one member of the two sequences with the highest pairwise sequence identity.

Levels of sequence homology between proteins of the training set and those for which predictions are analyzed here are even lower than within the training set; the highest pairwise sequence identity is 16% (*Rhodopseudomonas blastica* porin/Omp32) and the average is 11%. Given that no high-resolution structural information about the proteins with predicted topologies is available, and existing data are from a variety of experimental sources, the correspondence of predicted and experimentally derived topologies is good. This indicates that general features of OM protein topology, also applicable to proteins dissimilar to the training set, are modelled by the NN.

For the problem of OM protein topology prediction, the NN described here thus appears to be useful even though relatively little topological information about this class of proteins is available. With more structural information, derived from ongoing X-ray analyses of OM proteins in our and other groups, we believe that improvements in the accuracy of prediction will be possible. First, a higher number of training cases would marginally improve the results for the architecture of the NN described here (Equation 4). A higher gain in accuracy, however, is likely to be obtained by altering the architecture of the NN to make efficient use of the higher number of training cases then available, e.g., implementation of a larger hidden layer (Chandonia & Karplus, 1996).

Compared to methods based on physicochemical parameters of amino acids, an NN of a given architecture has the disadvantage that, after the training stage, the available information about properties of the system under study is not readily accessible, being encoded into a large set of numbers, the weights $\{w\}$. Although an a posteriori analysis of the weights may be possible in principle, there is no procedure available for deriving from the weights a set of rules with parameters for single amino acids and their interactions. This hinders a simple abstraction of the optimized weights into a scheme for human understanding of the rules that govern the system under study. Work is underway to probe our NN with random (computer-generated) sequences that might allow identification of common features leading to particularly high or low predicted z-values.

An NN, although seemingly complex at first sight, is a powerful and simple implementation of a general method adapting itself to a wide range of biological (and other) problems. As the output of an NN depends on the input in a nonlinear way (Equation 1) and the hidden layer allows for interactions among the input units, the NN training procedure may satisfactorily model complex biological systems that are hard to capture with simple rules.

Modeling of OM protein topology from the amino acid sequence alone has thus far relied on prediction of β-strands; in the absence of experimental information, the amino- and carboxy termini were often assumed to be in the periplasm. The topology was then derived from the β-strand prediction by connecting the strands in an antiparallel fashion. This approach appears justified at first sight as the known OM protein structures are particularly simple 16-stranded antiparallel β-barrels with nearest-neighbor connections between strands. Although simple, this topology of TM β-strand proteins gives rise to a particular difficulty with respect to constructing their topology from a given β-strand prediction: as adjacent strands are antiparallel, the direction of membrane of a given strand is implied by its position within the sequence of all strands. Thus, if a strand is missed by the prediction or if a strand is predicted where none exists, the resulting topology will be reversed starting at the point where the error occurred. The NN described here avoids this pitfall since

at each position of the sequence a local property (z-value) directly related to topology is predicted, independent of preceding stretches of amino acids. Errors in the prediction are not propagated to the sequence that follows, as both the location (periplasmic, TM, extracellular) and the direction of traversal of this region are derived from the z-values of the Cα-coordinates.

Even at the current, somewhat modest state of knowledge, our NN generates models of OM proteins that can be directly correlated with existing experimental data. We envisage our method to be used by molecular biologists working with OM proteins of gram-negative bacteria to assist in experiment design, e.g., introduction of His-tags, protease cleavage sites, or construction of fusion proteins.

No reliable tertiary structure (3D) prediction methods for OM proteins are currently available. Following Jones (1997), our method can be termed a 2D prediction method, as it goes beyond the information available from secondary structure prediction (1D) methods. The NN topology prediction is accessible as an online service on the World Wide Web (http://strucbio.biologie.uni-konstanz.de/~kay/om_topo_predict.html).

Materials and methods

Preparation of training sets for NN training

Porin structures solved by X-ray crystallography were obtained from the Protein Data Bank for *R. capsulatus* (2POR), *Rhodopseudomonas blastica* (1PRN), OmpF (2OMF), PhoE (1PHO), and maltoporin (1MAL). This set of publicly available structures was augmented by the locally solved structures of nonspecific porin from *Paracoccus* (Hirsch et al., 1997) and sucrose-specific porin (1A0S) from *S. typhimurium*. 2POR defined the reference frame, as the z-axis of its crystal structure is already aligned with the pore. Therefore, for 2POR, high z-values of the Cα coordinate imply residues of the extracellular loops whereas low z-values denote Cα positions in or near the periplasmic turns. The coordinate sets (2,388 cases altogether) of the other porins were structurally aligned with those of 2POR using SUPERIMPOSE (Diederichs, 1995). To obtain the target output values o_k, a transformation was obtained by normalizing the z-coordinates of Cα atoms from 2POR to lie in the 0.0–1.0 range. The same transformation was applied to all other Cα coordinates, yielding a total set of 2,388 cases. After normalization, z-values below 0.1 were truncated to 0.1 and z-values above 0.9 were truncated at 0.9, to avoid strong bias of the NN weights toward the extracellular loop or periplasmic turn regions.

Architecture of the NN

A backpropagation NN with 15 * 21 input units, two hidden units, and one output unit (Fig. 1A) was used for determining the weights and for prediction of (normalized) Cα z-values. The 20 amino acid types in a sliding window of 15 residues were used as input data, with an additional pseudo amino acid type indicating a position before the amino terminus, or beyond the carboxy terminus. This pseudo amino acid type is only required if the central position of the sliding window is between residues 1 and 7 inclusively, or between residues N-6 and N inclusively of the amino acid sequence (N being the total number of residues of the sequence). Each amino acid type was coded as a string of 21 "0" or "1" values (e.g., "10000 0000000000000000" for Ala, "01000000000000000000" for Cys, and so on), therefore requiring 15 * 21 input units. This NN topology is similar to that used by others (Holley & Karplus, 1989) for the prediction of secondary structure properties of soluble proteins. Compared to these studies our NN employs a somewhat smaller input window and only one output unit. The number of input amino acid positions used (15) encompasses at least one and often two β-strands. This window size was therefore judged large enough to represent essential sequence information about the amino acids at and near a given Cα position. One output unit, representing the normalized Cα coordinate, is the natural choice in the case of a smoothly varying and normalized target value, as no threshold parameter for the choice between two binary output values is required.

Cross-validation of training success

A NN can be used to abstract properties of input data that generally lead to preferred values of output data. However, a successful generalization of underlying principles in the prediction of properties by an NN requires a much larger number of cases than the number of adjustable parameters (weights). With the given architecture of the NN used here, there are $(15 * 21 + 1) * 2 + (2 + 1) * 1 = 635$ weights to be adjusted during the training phase. As the number of cases is less than a factor of four higher than the number of weights, we monitored the training success by cross-validation (Efron & Tibshirani, 1991) as a function of case number. For this purpose, the 2,388 available cases were randomly assigned to either a "training set" or a "test set." Cases from the training set were used for adjusting the weights and cases from the test set were used afterward to assess the quality of the prediction. The relative size of the training set with respect to the test set was varied to investigate the influence of case number in the training set and to obtain an estimate of the asymptotic training behavior.

It was shown (Qian & Sejnowski, 1988) that high sequence homology between training and test sets may lead to false indications of greater accuracy of NN performance. For the data under study, there are no homologous repeats within the sequences of the proteins used, and there is only one case of high pairwise sequence identity (60% for OmpF/PhoE). All other pairwise alignments show low (37% for *Blastica/Paracoccus*) or insignificant (<23%) sequence identity. To test the effect of homology on cross-validation results, the above calculations were repeated with PhoE deleted from the cases available, resulting in 2,058 cases instead of 2,388.

Note added in proof

After submission of this paper, we solved the X-ray structure of FhuA. Consistent with the NN prediction presented here, the architecture of the protein is dominated by a 22-stranded beta barrel. In addition, the first 160 residues form a "cork" domain which fills the central cavity of the barrel. This domain displays four additional beta-strands, several helices, loops, and coils.

Acknowledgments

We thank M. Ehrmann and W. Welte for discussion and comments on the manuscript and J.W. Coulton, A. Ferguson, and G. Moeck for critical reading of the manuscript. J. Breed thanks EMBO for a Long-Term Fellowship (ALTF751-1995).

References

Buchanan S, Smith B, Venkatramani L, vd Helm D, Deisenhofer J. 1996. Oral Presentation Ms04.03a.08 at IUCr XVII Congress and General Assembly. Seattle, WA.

Chandonia J-M, Karplus M. 1996. The importance of larger data sets for protein secondary structure prediction with neural networks. *Protein Sci 5*:768–774.

Chen R, Schmidmayr W, Kramer C, Chen-Schmeisser U, Henning U. 1980. Primary structure of major outer membrane protein II (*ompA* protein) of *Escherichia coli* K-12. *Proc Natl Acad Sci USA 77*:4592–4596.

Cowan SW, Rosenbusch JP. 1994. Folding pattern diversity of integral membrane proteins. *Science 264*:914–916.

Deisenhofer J, Epp O, Miki K, Huber R, Michel H. 1985. Structure of the protein subunits in the photosynthetic reaction center of *Rhodopseudomonas viridis* at 3 Å resolution. *Nature 318*:618–624.

Diederichs K. 1995. Structural superposition of proteins with unknown alignment and detection of topological similarity using a six-dimensional search algorithm. *Proteins Struct Funct Genet 23*:187–195.

Dubchak I, Holbrook SR, Kim S-H. 1993. Prediction of protein folding class from amino acid composition. *Proteins Struct Funct Genet 16*:79–91.

Efron B, Tibshirani R. 1991. Statistical data analysis in the computer age. *Science 253*:390.

Ehrmann M, Bolek P, Mondigler M, Boyd D, Lange R. 1997. TNTIN and TNTAP: Mini-transposons for site-specific proteolysis. *Proc Natl Acad Sci 94*:13111–13115.

Ferguson A, Breed J, Diederichs K, Welte W, Coulton JW. 1998. An internal affinity-tag for purification and crystallization of the siderophore receptor FhuA, an integral outer membrane protein from *Escherichia coli* K-12. *Protein Sci 7*:1636–1638.

Freudl R. 1989. Insertion of peptides into cell-surface-exposed areas of *Escherichia coli* OmpA protein does not interfere with export and membrane assembly. *Gene 82*:229–236.

Freudl R, MacIntyre S, Degen M, Henning U. 1986. Cell surface exposure of the outer membrane protein OmpA of *Escherichia coli* K-12. *J Mol Biol 188*:491–494.

Georgiou G, Stephens DL, Stathopoulos C, Poetschke HL, Mendenhall J, Earhart CF. 1996. Display of β-lactamase on the *Escherichia coli* surface: Outer membrane phenotypes conferred by Lpp′-OmpA′-β-lactamase fusions. *Protein Eng 9*:239–247.

Geourjon G, Deléage G. 1994. SOPM: A self-optimized method for protein secondary structure prediction. *Protein Eng 7*:157–164.

Gerbl-Rieger S, Engelhardt H, Peters J, Kehl M, Lottspeich F, Baumeister W. 1992. Topology of the anion-selective porin Omp32 from comamonas acidovorans. *J Struct Biol 108*:14–24.

Gromiha MM, Majumdar R, Ponnuswamy PK. 1997. Identification of membrane spanning β strands in bacterial porins. *Protein Eng 10*:497–500.

Hirsch A, Breed J, Saxena K, Richter O-MH, Ludwig B, Diederichs K, Welte W. 1997. The structure of porin from *Paracoccus denitrificans* at 3.1 Å resolution. *FEBS Lett 404*:208–210.

Hirst JD, Sternberg MJE. 1991. Prediction of ATP-binding motifs: A comparison of a perceptron-type neural network and a consensus sequence method. *Protein Eng 6*:615–623.

Holley LH, Karplus M. 1989. Protein secondary structure prediction with a neural network. *Proc Natl Acad Sci USA 86*:152–156.

Jähnig F. 1990. Structure predictions of membrane proteins are not that bad. *Trends Biochem Sci 15*:93–95.

Jones DT. 1997. Progress in protein structure prediction. *Curr Op Str Biol 7*:377–387.

Klug CS, Su WY, Feix JB. 1997. Mapping of the residues involved in a proposed beta-strand located in the ferric enterobactin receptor FepA using site-directed spin-labeling. *Biochemistry 36*:12027–13033.

Kneller DG, Cohen FE, Langridge R. 1990. Improvements in protein secondary structure prediction by an enhanced neural network. *J Mol Biol 214*:171–182.

Koebnik R. 1996. *In vivo* membrane assembly of split variants of the *E. coli* outer membrane protein OmpA. *EMBO J 15*:3529–3537.

Koebnik R, Braun V. 1993. Insertion derivatives containing segments of up to 16 amino acids identify surface- and periplasm-exposed regions of the FhuA outer membrane receptor of *Escherichia coli* K-12. *J Bacteriol 175*:826–839.

Koebnik R, Krämer L. 1995. Membrane assembly of circularly permuted variants of the *E. coli* outer membrane protein OmpA. *J Mol Biol 250*:617–626.

Liu J, Rutz JM, Klebba PE, Feix JB. 1994. A site-directed spin-labelling study of ligand-induced conformational change in the ferric enterobactin receptor, FepA. *Biochemistry 33*:13274–13283.

Milik M, Kolinski A, Skolnick J. 1995. Neural network system for the evaluation of side-chain packing in protein structures. *Protein Eng 8*:225–236.

Moeck GS, Bazzaz BSF, Gras MF, Ravi TS, Ratcliffe MJH, Coulton JW. 1994. Genetic insertion and exposure of a reporter epitope in the ferrichrome-iron receptor of *Escherichia coli* K-12. *J Bacteriol 176*:4250–4259.

Moeck GS, Ratcliffe MJH, Coulton JW. 1995. Topological analysis of the *Escherichia coli* ferrichrome-iron receptor by using monoclonal antibodies. *J Bacteriol 177*:6118–6125.

Morona R, Klose M, Henning U. 1984. *Escherichia coli* K-12 outer membrane protein (OmpA) as a bacteriophage receptor: Analysis of mutant genes expressing altered proteins. *J Bacteriol 159*:570–578.

Morona R, Kramer C, Henning U. 1985. Bacteriophage receptor area of outer membrane protein OmpA of *Escherichia coli* K-12. *J Bacteriol 164*:539–543.

Murphy CK, Kalve VI, Klebba PE. 1990. Surface topology of the *Escherichia coli* K-12 ferric enterobactin receptor. *J Bacteriol 172*:2736–2746.

Muskal SM, Kim SH. 1992. Predicting protein secondary structure content: A tandem neural network approach. *J Mol Biol 225*:713–727.

Newton SMC, Allen JS, Cao Z, Qi Z, Jiang X, Sprencel C, Igo JD, Foster SB, Payne MA, Klebba PE. 1997. Double mutagenesis of a positive charge cluster in the ligand-binding site of the ferric enterobactin receptor, Fep A. *Proc Natl Acad Sci USA 94*:4560–4565.

Ostermeier C, Iwata S, Michel H. 1996. Cytochrome *c* oxidase. *Curr Op Struct Biol 6*:460–466.

Paul C, Rosenbusch JP. 1985. Folding patterns of porin and bacteriorhodopsin. *EMBO J 4*:1593–1597.

Presnell SR, Cohen FE. 1993. Artificial neural networks for pattern recognition in biochemical sequences. *Ann Rev Biophys Biomol Struct 22*:283–298.

Qian N, Sejnowski TJ. 1988. Predicting the secondary structure of globular proteins using neural network models. *J Mol Biol 262*:865–884.

Ried G, Koebnik R, Hindennach I, Mutschler B, Henning U. 1994. Membrane topology and assembly of the outer membrane protein OmpA of *Escherichia coli* K-12. *Mol Gen Genet 243*:127–135.

Rost B, Fariselli P, Casadio R, Sander C. 1995. Prediction of helical transmembrane segments at 95% accuracy. *Protein Sci 4*:521–533.

Rost B, Sander C. 1994a. Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins Struct Funct Genet 19*:55–72.

Rost B, Sander C. 1994b. Conservation and prediction of solvent accessibility in protein families. *Proteins Struct Funct Genet 20*:216–226.

Rumelhart DE, Hinton GE, Williams RJ. 1986. Learning representations by back-propagating errors. *Nature 323*:533–536.

Ruppert A, Arnold N, Hobom G. 1994. OmpA-FMDV VP1 fusion proteins: Production, cell surface exposure and immune responses to the major antigenic domain of foot-and-mouth disease virus. *Vaccine 12*:492–498.

Schertler GFX, Villa C, Henderson R. 1993. Projection structure of rhodopsin. *Nature 362*:770–772.

Schirmer T, Cowan SW. 1993. Prediction of membrane-spanning beta-strands and its application to maltoporin. *Protein Sci 2*:1361–1363.

Sonntag I, Schwarg H, Hirota Y, Henning U. 1978. Cell envelope and shape of *Escherichia coli*: Multiple mutants missing the outer membrane lipoprotein and other major outer membrane proteins. *J Bacteriol 136*:280–285

Srikumar R, Chin AC, Vachon V, Richardson CD, Ratcliffe MJH, Saarinen L, Käyhty H, Mäkelä PH, Coulton JW. 1992a. Monoclonal antibodies specific to porin of *Haemophilus influenzae* type b: Localization of their cognate epitopes and tests of their biological activities. *Mol Microbiol 6*:665–676.

Srikumar R, Dahan D, Gras MF, Ratcliffe MJH, van Alphen L, Coulton JW. 1992b. Antigenic sites on porin of *Haemophilus influenzae* type b: Mapping with synthetic peptides and evaluation of structure predictions. *J Bacteriol 174*:4007–4016.

Srikumar R, Dahan D, Arhin FF, Tawa P, Diederichs K, Coulton JW. 1997. Porins of *Haemophilus influenzae* type b mutated in loop 3 and in loop 4. *J Biol Chem 272*:13614–13621.

Stathopoulos G. 1996. An alternative topological model for *Escherichia coli* OmpA. *Protein Sci 5*:170–173

Tommassen J. 1988. Biogenesis and membrane topology of outer membrane proteins in *Escherichia coli*. In: Op den Kamp JAF, ed. *Membrane biogenesis. NATO ASI Series Vol. H16*. Berlin, Heidelberg: Springer.

Unger VM, Hargrave PA, Baldwin JM, Schertler GFX. 1997. Arrangement of rhodopsin transmembrane α-helices. *Nature 389*:203–206.

Vogel H, Jähnig F. 1986. Models for the structure of outer-membrane proteins of *Escherichia coli* derived from Raman spectroscopy and prediction methods. *J Mol Biol 190*:191–199.

Welte W, Weiss MS, Nestel U, Weckesser J, Schiltz E, Schulz GE. 1991. Prediction of the general structure of OmpF and PhoE from the sequence and structure of porin from *Rhodobacter capsulatus*. Orientation of porin in the membrane. *Biochim Biophys Acta 1080*:271–274.

Yoshikawa S. 1997. Beef heart cytochrome *c* oxidase. *Curr Op Struct Biol 7*:574–579.

Zeth K, Schnaible V, Przybylski M, Welte W, Diederichs K, Engelhardt H. 1998. Crystallization and mass spectrometric analysis of native and chemically modified porin Omp32 from Comamonas acidovorans. *Acta Cryst D54*:650–653.