Eur Biophys J (1996) 24: 165–178

European Biophysics Journal

A predictor of transmembrane α-helix domains of proteins based on neural networks

Rita Casadio, Piero Fariselli, Chiara Taroni, Mario Compiani*

Laboratory of Biophysics, Department of Biology, University of Bologna, Via Irnerio 42, I-40126 Bologna, Italy
(Tel.: 0039-51-351284, Fax: 0039-51-242576, e-mail: G4XBO3B1@CINE88.CINECA.IT)

Received: 4 November 1994 / Accepted in revised form: 15 September 1995

Abstract. Back-propagation, feed-forward neural networks are used to predict α-helical transmembrane segments of proteins. The networks are trained on the few membrane proteins whose transmembrane α-helix domains are known to atomic or nearly atomic resolution. When testing is performed with a jackknife procedure on the proteins of the training set, the fraction of total correct assignments is as high as 0.87, with an average length for the transmembrane segments of 20 residues. The method correctly fails to predict any transmembrane domain for porin, whose transmembrane segments are β-sheets. When tested on globular proteins, lower and upper limits of 1.6 and 3.5% for a total of 26826 residues are determined for the mispredicted cases, indicating that the predictor is highly specific for α-helical domains of membrane proteins. The predictor is also tested on 37 membrane proteins whose transmembrane topology is partially known. The overall accuracy is 0.90, two percentage points higher than that obtained with statistical methods. The reliability of the prediction is 100% for 60% of the total 18242 predicted residues of membrane proteins. Our results show that the local directional information automatically extracted by the neural networks during the training phase plays a key role in determining the accuracy of the prediction.

Key words: Membrane proteins – Prediction of transmembrane α-helices – Protein folding – Protein structure prediction – Pattern recognition – Artificial neural networks

Introduction

Membrane proteins are a relevant class of complex biomolecular systems responsible of several important functions such as signal recognition, transport phenomena, energy translocation and conservation in the living cell (Von Heijne 1988; Jennings 1989; Traxler et al. 1993). A major goal in protein biophysics is the prediction of their transmembrane location and topology.

Presently only few membrane proteins are resolved at atomic or nearly atomic resolution, including photosynthetic reaction centres (Deisenhofer et al. 1985; Feher et al. 1989), bacteriorhodopsin (Henderson et al. 1990), the plant light-harvesting complex (Kühlbrandt et al. 1994) and porin (Weiss et al. 1991). This is in contrast with the large number of primary sequences known for this protein class (as well as for globular proteins), that is exponentially growing as a consequence of the widespread application of recombinant DNA techniques in molecular biology (Bowie et al. 1991; Chothia 1992).

Integral membrane proteins interact with the lipid bilayer to a variable extent and can be monotopic, bitopic and polytopic, depending on the number of times the polypeptide spans the membrane (Jennings 1989). The surface of membrane proteins is hydrophobic in the region in contact with the alkane chains of the lipids and polar in the region in contact with the acqueous phases and/or the interfaces. As a result, the crystallization of membrane proteins is particularly difficult as compared to that of globular proteins (Michel 1983).

The transmembrane segments of crystallized membrane proteins are typically apolar helices, perpendicular or partially tilted with respect to the normal to the membrane bilayer (Michel et al. 1986). The only exception so far described at atomic resolution is the structure of the bacterial outer membrane porin, where the transmembrane segments are β-strands arranged in a 16-stranded β-barrel (Weiss et al. 1991).

The topography and topology of membrane proteins is presently determined with experimental approaches, such as gene fusion analysis and biochemical probes, and/or theoretical methods, essentially aimed at the recognition of the hydrophobic membrane spanning stretches distributed throughout the aminoacid sequence (Traxler et al. 1993). Each of these domains, together with its flanking

* *Present address:* Department of Chemical Sciences, University of Camerino, Camerino, Italy

Correspondence to: R. Casadio

hydrophilic regions, constitutes a topological determinant, that can insert independently into the membrane (Popot and De Vitry 1990). The most evident of the topogenic signals so far described (Traxler et al. 1993) is the prevalence of positively charged residues in the cytoplasmic loops of the bacterial inner membrane proteins (Von Heijne 1992).

The theoretical search for topological determinants along the protein sequence is usually performed by evaluating a running-average hydrophobicity over a sliding window of appropriate length on the basis of hydrophobicity scales (Kyte and Doolittle 1982; Engelman et al. 1986) and/or statistical propensities for each residue to be buried in the membrane phase (Kuhn and Leigh 1985; Klein et al. 1985; Rao and Argos 1986). Several hydrophobicity scales were calculated on the basis of different physicochemical properties of the aminoacid residues and/or their relative propensities to be found in a given conformational structure or frequency of occurrence in the known membrane-spanning regions (Schultz 1988; Fasman 1989; Fasman and Gilbert 1990; Degli Esposti et al. 1990). For amphipathic membrane structures, predictive accuracy was further improved by evaluating the periodicity of the hydrophobicity (Eisenberg et al. 1984; Cornette et al. 1987).

Most of these methods implicitly assume that the predicted segments are α-helices normal to the membrane, with a typical length of 17–25 residues and that these segments are associated with the peaks found in the hydrophobicity plot, that displays the average hydrophobicity of each residue versus its position along the protein sequence (Fasman and Gilbert 1990). The window length and the decision functions, crucial to the accuracy of the prediction, are however empirically tailored on specific sets of membrane proteins and are not generally valid.

A step forward in the statistical analysis of the parameters relevant for the transmembranicity analysis was made by calculating the window length and weight values for the optimal predictive accuracy directly on the reference chains of the reaction centers and bacteriorhodopsin (Edelman 1993). The mathematical procedures used were linear (Edelman and White 1989) and a more efficient quadratic minimization (Edelman 1993). This so called "optimal predictor" could then predict membrane proteins of different functional classes in agreement with the putative models: the error rate ranged from 6 to 14% depending on the minimum length chosen for the transmembrane segment.

A totally different approach takes advantage of the information contained in the over 100 proteins of documented transmembrane sequences extracted from the SWISS-PROT data base (Bairoch and Boeckmann 1992). The topogenic parameters for the residues were evaluated and used to recognize topological models of all-helical membrane proteins with expectation maximization of the compatibility of a given sequence with a given topology and secondary structure (Jones et al. 1994). Alternatively, the information deriving from the single sequence was extended with multiple sequence alignments of related proteins to re-evaluate the average membrane propensity values for each residue along the alignment (Persson and Argos 1994). In either case, the methods developed for prediction of transmembrane segments give high predictive accuracy as compared to the putative models of membrane proteins.

From the above considerations it appears that the role of the local residue sequence is crucial in determining the transmembrane segments. The prediction of membrane proteins can be therefore considered as a pattern recognition problem. In this case, a powerful approach to its solution can be the use of artificial neural networks (Lippmann 1987; Müller and Reinhardt 1990). Neural networks have been used for the prediction of secondary structures of globular and membrane proteins (Qian and Sejnowski 1988; Holley and Karplus 1989; Kneller et al. 1990; Hirst and Sternberg 1992; Presnell and Cohen 1993; Fariselli et al. 1993). When three structural types (α-helix, β-strand and random coil) are discriminated, the predictive accuracy ranks around 60–65%, and is similar to that obtained with other statistical methods (Fasman 1989; Garnier and Levin 1991). This limit can be exceeded by six to eight percentage points by using evolutionary information in the form of multiple sequence alignment (Rost and Sander 1993; Rost et al. 1994). Neural networks were also applied to several problems in nucleic acid recognition with a predictive accuracy exceeding 90% (Hirst and Sternberg 1992; Presnell and Cohen 1993).

In the present paper, we develop a predictor for determining the topography of membrane proteins within the framework of the artificial neural network approach. The assignments of transmembrane regions is performed on the basis of the directional information that the network automatically extracts during the learning phase from the data set of membrane proteins known at atomic resolution. It is shown both on statistical and on a protein basis that the accuracy of this predictor is higher than that of the predictive methods previously described.

2. Methods

The structural data base

Our data base comprises membrane and amphipathic proteins known at atomic (or nearly atomic) resolution (Table 1). These proteins were grouped in three different training sets, not including or including amphipathic α-helices (Table 2). Assignment of the secondary structure of each protein residue is done according to the models derived from structural data. The plant light-harvesting a/b-protein complex, recently resolved at 3.4 Å resolution by electron crystallography (Kühlbrandt et al. 1994), is included in the testing set of membrane proteins as a benchmark for our method.

The testing sets of membrane proteins comprise the β-strand rich, pore-forming protein porin (POR) from the outer membrane of *Rhodobacter capsulatus* (Weiss et al. 1991) and thirty seven membrane proteins purposely selected in order to include the most relevant functional classes of membrane proteins (see legend to Table 2). The same set (TM_{37}) was also used by Edelman (1993) to test his optimal predictor based on quadratic minimization.

Table 3 illustrates the composition of the testing sets of globular proteins according to the Brookhaven code

Table 1. The training set of transmembrane α-helices

Symbol[a]	Residues	α (%)	α_T (%)	N_T
RCVH	258	18	10	1
RCVL	273	65	50	5
RCVM	323	64	43	5
RCSH	260	17	10	1
RCSL	281	66	42	5
RCSM	307	65	46	5
BRHO	248	n.d.	66	7
AMLT	26	81	81	1
ALAM	19	100	100	1
DLYS	26	100	100	1

[a] RCVH, RCVL, RCVM are the high, light and medium weight subunits of the photosynthetic reaction center of *Rhodopseudomonas viridis* (Deisenhofer et al. 1985); RCSH, RCSL and RCSM are the corresponding subunits of the photosynthetic reaction center of *Rhodobacter sphaeroides* (Feher et al. 1989). BRHO, AMLT, ALAM and DLYS stand for bacteriorhodopsin (Henderson et al. 1990), melittin (Terwillinger et al. 1982), alamethicin (Fox and Richards 1982) and δ-haemolysin (Mellor et al. 1988), respectively. α (%) = percentage of α-helix structure; α_T (%) = percentage of transmembrane α-helices; N_T = number of transmembrane α-helix segments

Table 2. Training and testing sets of membrane proteins

Symbol	Residues	Residues$^{\circ}_{T}$	N°_{T}
L_{NET1}	1951	749	29
L_{NET2}	1977	770	30
L_{NET3}	2022	815	32
T_{37}	18242	4469[a]	201[a]

The composition of the different training (L) and testing (T) sets of membrane proteins is listed below:

L_{NET1} = {RCVH, RCVL, RCVM, RCSH, RCSL, RCSM, BRHO}
$L_{NET2} = L_{NET1} \cup$ {AMLT}
$L_{NET3} = L_{NET2} \cup$ {ALAM, DLYS}

T_{37} = {β_2 adrenergic receptor (human) (Strosberg 1991), m_1 acetylcholine receptor (human) (Strosberg 1991), rhodopsin (bovine) (Khorana 1992), glycine receptor (rat, mature form) (Grenningloh et al. 1987), $GABA_A$ receptor (α subunit, human, mature form) (Schofield et al. 1987), nicotinic acetylcholine receptor (α, β, γ and δ subunits, *Torpedo californica*, mature forms) (Noda et al. 1983 a and 1983 b), K^+ channel (rat brain cortex, RCK1 and RCK5) (Stumher et al. 1989), Ca^{2+} channel (L type, α_1 subunits, rabbit muscle) (Tanabe et al. 1987), Na^+ channel (*Electrophorus electricus* electroplax) (Kayano et al. 1988), connexins (*Xenopus laevis* cx30 and cx38, rat liver cx32 and cx26, rat heart cx43) (Gimlich et al. 1988), Ca^{2+}-ATPase (sarcoplasmic reticulum, rabbit fast twitch muscle) (Brandl et al. 1986), Na^+-K^+-ATPase (α and β subunits, sheep kidney, mature forms) (Shull et al. 1986), light-harvesting complex II (pea, mature form) (Burgi et al. 1987), B870 (α and β subunits, *Rhodobacter sphaeroides*, mature form) (Brunisholtz et al. 1986), photosystem II (D1 subunit spinach) (Sayre et al. 1986), cytochrome P450IIB1 (rat) (Fujii-Kuriyama et al. 1982), cytochrome b_5 (rabbit) (Takagaki et al. 1983), NADPH-P450 reductase (rat) (Black and Coon 1982), NADH-cytochrome b_5 reductase (Ozols et al. 1984), lactose permease (*Escherichia coli*) (Foster et al. 1983), MALF (*Escherichia coli*) (McGovern et al. 1991), MotB (*Escherichia coli*) (Stader et al. 1986), Tsr (serine chemoreceptor, *Escherichia coli*) (Manoil and Beckwith 1986), Lep (leader peptidase I, *Escherichia coli*) (Moore and Miura 1987), lipophilin (myelin proteolipid, human, mature form) (Popot et al. 1991), glycophorin A (human, mature form) (Ross et al. 1982), band 3 anion exchanger (mouse) (Kopito and Lodish 1985)}

° Transmembrane residues and number of transmembrane segments (N_T)
[a] The values are putative for T_{37}

(Bernstein et al. 1977). Residue assignment of a given type of secondary structure is done with the define secondary structure protein program (Kabsch and Sander 1983). Most of the globular proteins known at atomic resolution are grouped in four different sets. G_{15} is essentially the same set used by Edelman to test its predictor on globular proteins and to compare it with other predictive methods (Edelman 1993). The two other sets of globular proteins, termed G_{28} and G^*_{28}, include the most representative proteins previously classified all-α (G_{28}) and all-β (G^*_{28}) (Chothia and Finkelstein 1990). The largest set (G_{86}) comprises proteins of the ($\alpha+\beta$), α/β and none classes ($G_{28} \cap G^*_{28} \cap G_{86} = \emptyset$).

The homology between all pair of proteins of the training and testing sets is determined by applying the FASTP program for comparison of sequences (Lipman and Pearson 1985). The z score (the number of standard deviations from the mean alignment score for the sequences of similar lengths) is in all cases very low ($z \leq 6$) (Lipman and Pearson 1985). The only exceptions are the significant homology existing between the L and M subunits of the reaction centre and between the similar subunits of the reaction centres from different sources ($z > 130$ and $z > 30$, respectively).

The network architecture

The most efficient model for our prediction task consists of two networks (Fig. 1). The first one (basic network in Fig. 1) is a perceptron without hidden layers, performing a supervised learning phase. The signals fom the input units (neurons) are multiplied by the vector of the connections (weights) and fed forward to the non-linear processing output node, which is triggered by a sigmoid function (Fig. 1). Optimal mapping of input to output patterns is automatically learned by the network by means of the backpropagation algorithm (Rumelhart et al. 1986). Steepest descent (with an appropriate learning rate) is used to compute the changes that must be applied to the connections and thresholds in order to minimize the error function, which quantifies the difference between the actual and the desired outputs of the network. The classification of each residue is carried out by considering a distinguishing value of 0.5 to discriminate between transmembrane and non-transmembrane classes.

Sequences of outputs from the first network are used as inputs to the second network (cascaded network in Fig. 1) which, as is detailed below, has a filtering effect. The cascaded network performs best with one hidden layer containing two hidden nodes. The hidden neurons enhance the computational power of the network and accelerate the learning process.

Table 3. Testing sets of globular proteins

Symbol	Residues	α (%)	β (%)
G_{15}	3362	38	16
G_{28}	4101	60	1
G^{*}_{28}	4990	4	42
G_{86}	17735	28	19

The testing sets of globular proteins (G) contained different numbers of proteins as indicated by the number-index. Their protein composition is listed below, following the Brookhaven codes:

G_{15} = {1ABP, 1AZU, 1BP2, 1CYT, 1MBN, 2APR, 2CCP, 2LZM, 2SNS, 3ADK, 3B5C, 3GRS, 3TLN, 5CPA, 5CPV}

G_{28} = {1BP2, 1CC5, 1CCR, 1CTS(A), 1CYC, 1ECA, 1ECD, 1FDH(A,G), 1HDS(A, B), 1MBC, 1MBD, 1P2P, 2CCY(A), 2CYP, 2DHB(A,B), 2HMQ(A), 2LH1, 2LHB, 2LZM, 3C2C, 3CPV, 3HHB(A,B), 3ICB, 3WRP}

G^{*}_{28} = {1ACX, 1AZU, 1FC2(C), 1GCR, 1MCP(H,L), 1PCY, 1PFC, 1REI(A), 1TGS(Z), 2ALP, 2APP, 2APR, 2CNA, 2GCH(3), 2IG2(H,L), 2KAI(A,B), 2PAB(A), 2RHE, 2SGA, 2SOD(O), 2TBV(A), 3EST, 3RP2(A), 3SGB(I), 4APE}

G_{86} = {1ABP, 1CTF, 1CTX, 1CY3, 1ETU, 1FX1, 1FXB, 1GCN, 1GOX, 1GP1(A), 1HIP, 1HOE, 1LLC, 1MEV, 1MON(A), 1NXB, 1PPT, 1PRC(C), 1PYP, 1RDG, 1RHD, 1RN3, 1RNT, 1SN3, 1TIM(A), 1UBQ, 1UTG, 1WSY(A,B), 2AAT, 2ACT, 2ATC(A,B), 2AZA(A), 2B5C, 2CAB, 2CDV, 2CHA(A), 2CPP, 2CRO, 2GN5, 2INS(A), 2LBP, 2LDX, 2LIV, 2PAZ, 2PFK(A), 2PKA(A,B), 2PLV, 2PRK, 2RSP(A), 2SBT(1,2), 2SNS, 2STV, 2TAA(A), 2TMV(P), 2TS1, 3ADK, 3BCL, 3CLN, 3FXC, 3GPD(R), 3GRS, 3HMG(A,B), 3HVP, 3PGK, 3PGM, 3TLN, 451C, 4CAT(A), 4DFR(A), 4FD1, 4FXN, 4MDH(A,B), 4SBV(A), 5CPA, 5TNC, 6LDH, 6LYZ, 8ADH, 9PAP, 9WGA(A)}

According to the network approach the adjustable parameters are: the number of training cycles, the size of the input window, the learning rate, the initial values of the connections, the number of hidden layers and the number of nodes in the hidden layers.

A binary encoding scheme is used for the input patterns of the basic network. It consists of groups with 20 units, each representing one of the amino acid residues, with as many different groups as the number of residues accommodated in the input window. The input window size is optimized as described in the Results section. The number of nodes in the output layer is routinely fixed at one.

The initial values of the connections between nodes are randomly chosen in the range $[-10^{-2}, 10^{-2}]$. Different initial sets of random values lead to negligible changes on the network performance.

The networks are implemented on a personal computer in C language.

Measures of accuracy

A number of quality indexes are currently available for comparing different predictive methods tested on known models (Schulz and Schirmer 1979). We focus on the ones listed below in order to evaluate the efficiency of our method and to compare it with other predictive algorithms.

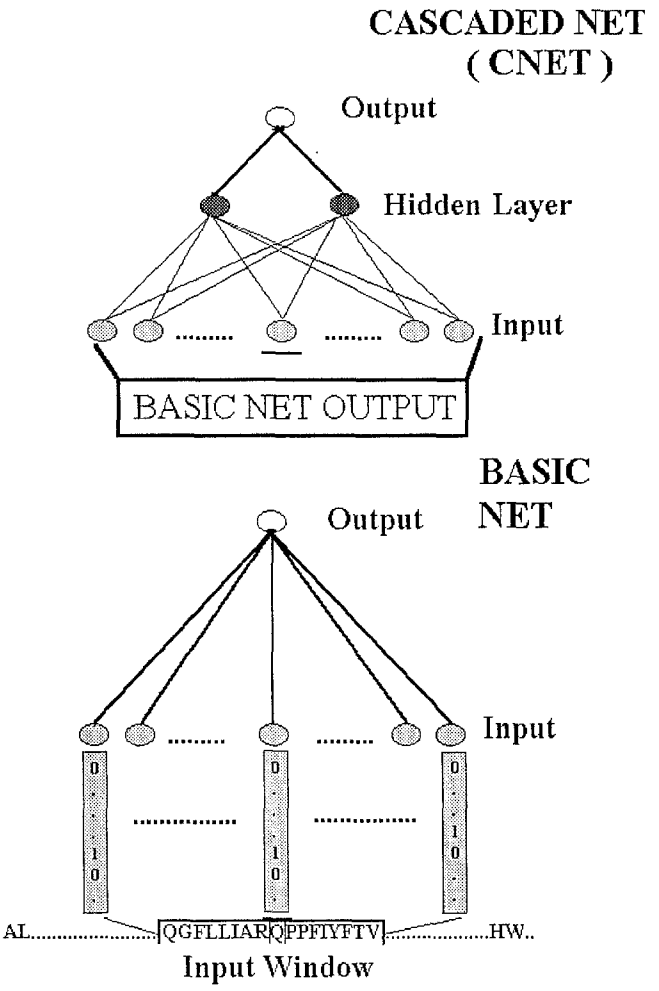

Fig. 1. Architecture of the neural networks used to predict the transmembrane domains of proteins. Grey, filled and open circles represent the neurons in the input, hidden and output layers, respectively. Lines between neurons in different layers indicate the connecting weights (w_{ij}). The residue sequence is encoded in the sliding input window with a binary scheme (bar symbols). Outputs of the simplest basic neural network without hidden layers are inputs to the cascaded network with two neurons in the hidden layer. At each neuron i, the activation is computed as:

$$a_i = \Sigma w_{ij}\, o_k + \vartheta_i$$

where o_k is the output of unit k and ϑ_i is a bias term (threshold value). Its output is computed according to a sigmoid trigger function:

$$o_i = [1 + \exp(-a_i)]^{-1}$$

Both networks are performing with a single output unit

The simplest measure of the predictive performance is given by the fraction of total correct predictions (commonly expressed on a percentage basis):

$$Q_3 = \Sigma_i P_i / N \quad (1)$$

where N is the total number of observed residues and P_i is the total number of residues correctly assigned to class i. The index is evaluated both on a residue and on a protein basis. In the latter case the associated standard deviation is also computed. Two classes are discriminated in the present study: transmembrane (T) and non-transmembrane (NT). Accordingly, two indexes are used, Q_T and Q_{NT}, in-

dicating the transmembrane and the non-transmembrane propensities, respectively:

$$Q_i = P_i/N_i = P_i/(P_i + U_i) \quad (2)$$

where N_i is the number of residues in the i-th class (i=T, NT), P_i is the number of residues correctly assigned to class i and U_i that of underpredicted cases.

As previously discussed (Fariselli et al. 1993), Q_3 and Q_i may be affected by the relative abundance of residues in either class. In this respect, a more meaningful measure of accuracy is the Matthew's correlation coefficient for the i-th class (C_i), which punishes over- and under-predictions (Matthews 1975). For the transmembrane class, it is:

$$C_T = (P_T R_T - U_T O_T)/[(R_T + O_T)(R_T + U_T)(P_T + U_T)(P_T + O_T)]^{1/2} \quad (3)$$

where P_T and R_T are the number of residues correctly assigned to and correctly rejected from class T, and U_T and O_T are the numbers of underpredicted and overpredicted cases. The coefficient ranges between –1 and 1, the latter being the value of the ideal correlation and 0 indicating a prediction no better than random. When only two classes are discriminated $C_T = C_{NT}$.

For evaluating the average length of a transmembrane segment we use the following index:

$$\langle L \rangle = \Sigma_i L_i / N_T \quad (4)$$

where L_i and N_T indicate the length of the i-th segment and the number of the transmembrane domains.

The accuracy of the predictor can be evaluated by computing the fractional overlap of the segments predicted by the network with those recorded in the data base. A measure of the fractional overlap of segments is the Sov index ($0 \leq Sov \leq 1$) (Rost et al. 1994):

$$Sov = \left[\Sigma_s (s_1 \cap s_2 / s_1 \cup s_2) L_{s_1}\right] / N \quad (5)$$

where s_1 and s_2 are the observed and predicted segments, respectively, and the summation is carried over all segment pairs; L_{s_1} is the length of the observed segment and N is the total number of residues. In the present work we evaluate the Sov index for the transmembrane (Sov_T) and non-transmembrane (Sov_{NT}) classes.

Finally, the reliability index (RI) is a measure of the likelihood of the correct prediction and is defined as (Rost and Sander 1993):

$$RI = INTEGER[10 \times (2 \times |o - 0.5|)] \quad (6)$$

where o is the network output; 0.5 is the distinguishing mark between a transmembrane and a non-transmembrane assignment (since the sigmoid trigger function permits maximal values of 1). The factors 10 and 2 allow to express RI with integer values from 0 to 9.

3. Results

Tuning the predictor

A major difficulty in applying neural networks to the predictive task of recognizing transmembrane segments of proteins is the paucity of examples known at atomic resolution (Table 2).

This prompted us to adopt a jackknife procedure, which consists in removing a protein from the training set, carrying out the learning phase on the remaining proteins and then predicting the protein which was removed. However, the jackknife method of testing may lead to higher scores for a test protein when it is significantly homologous to one or more proteins in the training set. This can occur especially when the training and testing phases are performed on sets of small dimensions, as in the present work. To overcome this difficulty, we optimize the basic network by focusing on the predictive efficiency of bacteriorhodopsin, which is not homologous to any other protein of the three different training sets ($z < 3$). The inclusion of amphipathic α-helices (such as melittin in L_{NET2} and melittin, δ-haemolysin and alamethicin in L_{NET3}), is done in order to provide the network with more examples of amphipathic patterns, considering that helix C of bacteriorhodopsin is the only amphipathic helix (Argos et al. 1982) among those in L_{NET1}.

The search in the parameter space of the network is performed at two different learning rates (0.1 and 0.01). At a fixed learning rate, the predictive accuracy of bacteriorhodopsin is clearly dependent on the size of the input window and on the number of training cycles (Fig. 2A, B). The three dimensional plots of the data indicate that the best prediction accuracy ($Q_3 = 0.73$) is obtained when the learning rate is 0.01 with a window size of 17 and a number of training cycles equal to 30. The values of the adjustable parameters of the network are not affected significantly either upon changing the protein which is removed or the composition of the training set (Table 2).

Four different indexes (see Measures of Accuracy) are used in Table 4 to assess the overall predictive performance of the perceptron (basic network) trained under the above conditions. The learning and generalization capabilities are measured by predicting the training sets without and with the jackknife method, respectively. On changing the composition of the training sets, the predictive efficiency is slightly affected and the highest score for the transmembrane assignment is obtained with NET2, trained on L_{NET2}. The score is however higher for the non-transmembrane than for the transmembrane class. This can be explained by considering that the residues of the training set (Table 2) are more abundant in the non-transmembrane class and that under our working conditions the prediction is statistically polarized towards the more abundant class (Compiani et al. 1992; Fariselli et al. 1993). The mean predicted length of the transmembrane segments in the training and testing sets is generally rather low as compared to the commonly accepted mean length of a transmembrane α-helix (Michel et al. 1986).

The cascaded network

The improvement provided by the introduction of the second network is apparent in Table 5. A relevant feature is the doubling of the average length of the transmembrane domains, in the range of the typical length of a transmem-

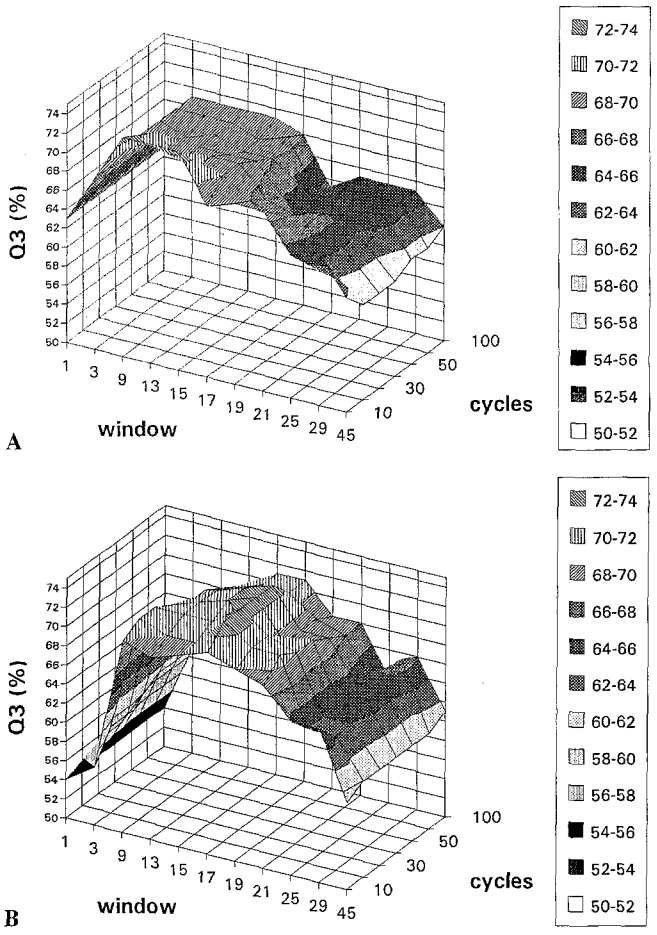

Fig. 2 A, B. Search in the parameter space of the network. The total efficiency for the prediction of bacteriorhodopsin obtained with NET2 is plotted as a function of the sliding window length and of the number of training cycles. Learning rates in the error function were fixed at 0.1 and 0.01 for the experiments shown in **A** and **B**, respectively

Table 4. Prediction with the basic networks

	Training				Testing[a]			
	Q_3	Q_T	Q_{NT}	$\langle L \rangle$	Q_{j3}	Q_{jT}	Q_{jNT}	$\langle L \rangle_j$
NET1	0.92	0.88	0.95	13.5	0.84	0.77	0.88	9.4
NET2	0.93	0.89	0.95	14.0	0.85	0.80	0.88	9.5
NET3	0.92	0.89	0.93	12.1	0.82	0.78	0.85	8.9

[a] The j index indicates that testing is performed with the jackknife procedure

Table 5. Prediction with the cascaded networks

	Training				Testing[a]			
	Q_3	Q_T	Q_{NT}	$\langle L \rangle$	Q_{j3}	Q_{jT}	Q_{jNT}	$\langle L \rangle_j$
CNET1	0.93	0.88	0.96	22.8	0.86	0.78	0.91	20.6
CNET2	0.93	0.88	0.96	22.9	0.87	0.79	0.92	19.0
CNET3	0.92	0.89	0.95	23.9	0.84	0.78	0.88	19.5

[a] As in the legend of Table 4

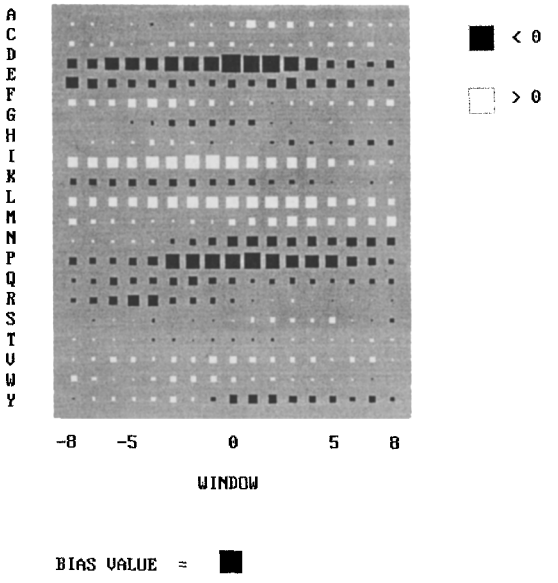

Fig. 3. Hinton diagram of the basic neural network showing the weights from the input units to the output unit when training is performed on the set of proteins L_{NET2} (see Table 2) with an input window of 17 residues. Residues are listed using the single letter code. White and blank squares represent positive (excitatory) and negative (inhibitory) weights, respectively. The area of each square is proportional to the value of the weight. The value of the bias term of the output unit is shown separately as an isolated square

brane α-helix perpendicular to the membrane plane (about 20 residues). The effect on the $\langle L \rangle$ index is evident for all the training sets. This result is in agreement with a sieving effect of the cascaded network which is expected to filter out spurious assignments, while summing up short neighboring fragments (Qian and Sejnowski 1988).

A search in the parameter space of the second network, while keeping fixed the first network, is obtained by training the network on the same set of training proteins as the first network and by predicting bacteriorhodopsin. This procedure gives rather flat surfaces as compared to that shown in Fig. 2 B. For this reason, changes in the adjustable parameters of the cascaded network do not impair significantly the improvement on the predictive efficiency of the augmented architecture. Routinely the optimal window size of the cascaded network is fixed at 17. Ten cycles of training and one hidden layer containing two nodes are sufficient to achieve the improvement on the overall prediction scores. On introducing the cascaded networks, it appears that the best generalization scores are obtained with CNET2 (Table 5). In the following, CNET1 is also used for comparison with other predictive methods.

The neural network based predictor

The contribution to the transmembrane class by each aminoacid residue is apparent in the Hinton diagram shown in Fig. 3, where the residues are represented with the single letter code and in alphabetical order. This diagram is a graphical representation of the values of the weights ob-

Table 6. Prediction of the training sets

Method	Q_3	Q_T	Q_{NT}	C_T	$\langle L \rangle$	Sov_T	Sov_{NT}	N_T
CNET1	0.93 (0.04)	0.86 (0.08)	0.96 (0.04)	0.81 (0.11)	22.4 (2.9)	0.81 (0.10)	0.90 (0.07)	31
CUT1/2	0.94 (0.04)	0.86 (0.08)	0.97 (0.04)	0.83 (0.08)	24.0 (2.5)	0.81 (0.10)	0.91 (0.08)	29
CNET2	0.92 (0.04)	0.87 (0.07)	0.96 (0.05)	0.80 (0.10)	22.0 (3.0)	0.82 (0.10)	0.87 (0.10)	32
CUT1/2	0.93 (0.04)	0.87 (0.07)	0.96 (0.05)	0.82 (0.08)	23.5 (2.9)	0.82 (0.10)	0.88 (0.10)	30
OP[a]	0.90 (0.05)	0.85 (0.05)	0.92 (0.08)	0.76 (0.11)	23.4 (3.4)	0.79 (0.05)	0.86 (0.12)	30

Values between brackets are standard deviations
[a] Edelman's optimal predictor based on quadratic minimization (Edelman 1993)

Table 7. Testing globular proteins

Method	G_{28}			G^*_{28}			G_{86}		
	$Q_3 (= Q_{NT})$	R_T	N_T	$Q_3 (= Q_{NT})$	R_T	N_T	$Q_3 (= Q_{NT})$	R_T	N_T
CNET1	0.94 (0.05)	281	37	0.93 (0.07)	408	48	0.94 (0.06)	1287	154
CUT1	1.00 (0.01)	19	1	0.98 (0.04)	132	6	0.99 (0.03)	276	15
CUT2	0.97 (0.04)	129	10	0.98 (0.05)	156	8	0.98 (0.04)	551	37
CNET2	0.93 (0.05)	301	40	0.93 (0.07)	440	51	0.93 (0.06)	1399	164
CUT1	1.00 (0.01)	20	1	0.98 (0.04)	134	6	0.99 (0.03)	319	17
CUT2	0.97 (0.04)	141	11	0.98 (0.05)	156	8	0.97 (0.05)	649	43

R_T and N_T are respectively the numbers of residues and segments assigned transmembrane

tained by training the optimized network NET2 on L_{NET2}. According to the network approach the strength of the inhibitory and/or excitatory signals for the transmembrane class of the residues depends on their relative position within the input window. Some residues such as aspartic and glutamic acids, proline and to a lesser extent glutamine exhibit a clear propensity to bias the output towards the non-transmembrane class. Others, such as cysteine, isoleucine, leucine, methionine, valine and tryptophan are excitatory for the transmembrane output. The remaining residues are either excitatory or inhibitory, depending on their position in the input window. The directional information, automatically transferred from the input to the output nodes by the network is then combined with the bias of the output unit, as described in the legend to Fig. 2. The general trend of the relative contribution of each residue to the Hinton diagram is not significantly modified upon changing the training set from L_{NET1} to L_{NET3} (data not shown).

An automatic cut-off procedure calibrated on the training sets

When, after the training phase, neural nets are used to predict the learning sets, a slight overprediction of transmembrane segments is noticed, as compared to the expected structures (31 and 32 instead of 29 and 30 for L_{NET1} and L_{NET2}, respectively). In order to correct for this, we introduce a criterion for rejecting false positive segments based on the logic AND function. On considering the prediction of the training set, a transmembrane segment is accepted provided that the length, height and area of the corresponding signal are ≥ than those of the smallest existing helix (CUT1). A less stringent version of the rejection criterion is obtained by requiring that the above values are > than those of the largest non-existing helix of the training set (CUT2). CUT2 can eventually also reveal transmembrane signals weaker than those accepted with CUT1.

The above filtering criterion is complemented by the following heuristic prescription, based on the notion that an upper bound for the length of transmembrane segments is known to exist (Edelman 1993). When the predicted segments exceed a fixed length of 36 residues (this happens three times when T_{37} is predicted), the stretch is automatically split in the neighborhood of the midpoint where the output signal has a local minimum.

As is shown in Table 6, on testing the training sets, CUT1 and CUT2 provide the same predictive accuracy (CUT1/2). The cutting procedure slightly improves the performance of the networks, as is evident on comparing the values of the accuracy indexes obtained before and after filtering. In Table 6, the learning capability of the networks on the training set is also compared to that of Edelman's optimal predictor (OP) which has been trained on the same set L_{NET1}. It is evident that after the cut-off procedure, neural networks can perform on the proteins of the training set better than OP.

Testing globular proteins and porin

In Table 7, the predictive capability of neural networks is tested on different sets of globular proteins in order to de-

termine the specificity of the predictor for membrane proteins. The predictors are trained specifically on hydrophobic α-helix transmembrane structures (Table 1). The aim of the first test on globular proteins is therefore to assess whether the features extracted by the predictor are mainly related to the motifs of secondary structure and/or to the hydrophobicity of the patterns. A possible way of grouping globular proteins is according to their structural anatomy, as previously described. The sets contain chains representative of the all-α (G_{28}), the all-β (G^*_{28}) and mixed α/β, ($\alpha+\beta$) and none classes (G_{86}). The data shown in Table 7 indicate that the predictors mispredict 6 to 7% of the residues of globular proteins, rather independently of the structural class. When CUT2 is applied, the false positive assignments decrease to 3.1–3.7% of the total predictions. With the most stringent CUT1, the percentage of false positives further decreases to 0.5% for the all-α proteins and to 1.6–1.8% for the G_{86} set, depending on the network used. For the all-β class the value of mispredictions is still 2.7%, suggesting that a possible dominant feature is the hydrophobicity of the patterns, which in globular proteins are mainly associated with β sheets. Interestingly, the rate of false positives predicted for G_{86} is mainly due to the α/β bundles included in this set.

Table 8. Testing globular proteins and porin

Method	G_{15}			POR		
	Q_3 (= Q_{NT})	R_T	N_T	Q_3 (= Q_{NT})	R_T	N_T
CNET1	0.95 (0.05)	193	27	0.95	14	3
CUT1	0.99 (0.01)	35	2	1.00	0	0
CUT2	0.99 (0.02)	46	3	1.00	0	0
CNET2	0.95 (0.05)	214	29	0.96	13	2
CUT1	0.99 (0.01)	36	2	1.00	0	0
CUT2	0.99 (0.02)	58	4	1.00	0	0
OP[a]	0.96 (0.05)	181	5–9[a]	0.99	4	0

[a] Edelman's optimal predictor (Edelman 1993); the lower and upper limits for N_T are obtained by considering different segment lengths (≥17 and ≥10, respectively)

For a direct comparison with other predictive methods, a further test is performed on G_{15}, a set comprising highly hydrophobic globular proteins of different structural classes. Porin, the only known β-strand containing membrane protein is also predicted. The results, shown in Table 8 indicate that the two neural networks respectively mispredict 5.7–6.4% of G_{15} and 4.7–4.3% residues of porin. CUT1 and CUT2 are, however, sufficient to decrease the mispredictions of G_{15} up to 1.0–1.4% and 1.1–1.7%, depending on the network used, and to completely suppress the false positives of porin.

For comparison, the prediction of G_{15} and porin with the Edelman's optimal predictor is also shown in Table 8. This method gives 5% of false positives and is comparable to that obtained with neural networks devoid of the cutting procedure.

Testing membrane proteins

Our membrane protein predictor was tested on T_{37}. Although the location of some of the transmembrane domains of the testing proteins contained in T_{37} is still controversial, models based on experimental results obtained with different approaches allow the definition of a likely topography (as can be found in the references listed in the legend of Table 2). The accuracy of the prediction of neural networks is therefore evaluated on the available models (Table 9) and compared to that of Edelman's optimal predictor (OP). Q_3 obtained with CNET1 (trained with the same training set as OP) is slightly better than the one obtained with OP. As expected on the basis of the results shown in Table 5, neural nets are underpredicting the transmembrane class. The higher score obtained on the transmembrane class by OP can be traced back to the higher rate of false positives as compared to neural networks when globular proteins are predicted (see Table 8). The putative number of 201 transmembrane segments in T_{37} (see Table 2) is within the upper and lower limits of the predicted transmembrane domains obtained with neural networks. More significantly, the values of the C_T and Sov_T indexes indicate that the overlap between putative and predicted segments is also quite satisfactory.

Table 9. Testing membrane proteins

Method	T_{37}							
	Q_3	Q_T	Q_{NT}	C_T	⟨L⟩	Sov_T	Sov_{NT}	N_T
CNET1	0.88 (0.05)	0.83 (0.12)	0.89 (0.08)	0.68 (0.14)	19.0 (6.4)	0.70 (0.10)	0.85 (0.09)	273
CUT1	0.90 (0.06)	0.77 (0.21)	0.92 (0.07)	0.70 (0.19)	23.3 (3.2)	0.66 (0.18)	0.84 (0.14)	185
CUT2	0.90 (0.06)	0.82 (0.13)	0.91 (0.08)	0.71 (0.14)	22.3 (3.6)	0.71 (0.12)	0.86 (0.10)	214
CNET2	0.88 (0.05)	0.84 (0.12)	0.88 (0.08)	0.67 (0.13)	19.1 (6.3)	0.69 (0.12)	0.85 (0.09)	278
CUT1	0.89 (0.06)	0.78 (0.21)	0.91 (0.08)	0.69 (0.19)	23.5 (3.1)	0.66 (0.18)	0.84 (0.13)	187
CUT2	0.89 (0.06)	0.83 (0.12)	0.91 (0.08)	0.71 (0.14)	22.5 (3.6)	0.71 (0.11)	0.86 (0.10)	218
OP[a]	0.88 (0.06)	0.90 (0.09)	0.87 (0.08)	0.70 (0.12)	25.5 (4.8)	0.70 (0.10)	0.84 (0.10)	229

[a] Edelman's optimal predictor (Edelman 1993)

In Fig. 4, at a given fixed value of the reliability index (Eq. 6), the percentage number of predicted residues in T_{37} and the corresponding total predictive accuracy evaluated on a residue basis are shown for the basic and the cascaded networks. It appears that the fraction of residues characterized by increasing values of the reliability index is progressively decreasing. Concomitantly, the predictive accuracy ranges from 87 to 100%; this indicates that even the less reliably predicted residues are not randomly classified. The cascaded network provides a percentage number of residues with a given reliability that is larger than that corresponding to the basic network. When the reliability limit is 0.7, 60% of the residues comprised in T_{37} are predicted with an accuracy of about 100%.

As an example, the performance of our method on the recently crystallized light-harvesting complex is shown in Fig. 5, where the net output (transmembranicity) is depicted along the residue sequence. In agreement with the putative transmembrane regions of this protein, three α-helical membrane segments are recognized with values of Q_3, C_T and Sov_T equal to 0.89, 0.72 and 0.63, respectively.

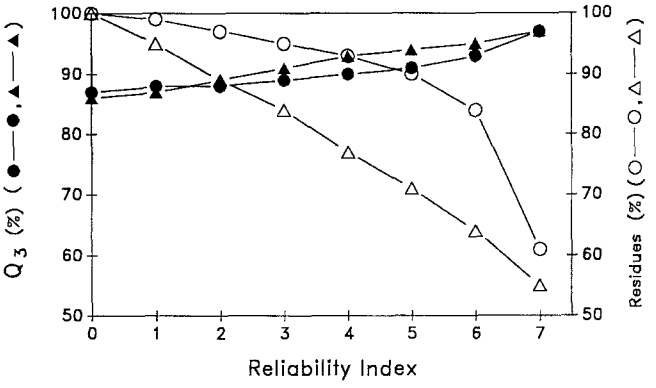

Fig. 4. The accuracy of the prediction for residues with a reliability index equal to a given value. Open and filled symbols represent respectively the percentage number of residues and the corresponding accuracy of the prediction at a given cut-off of the index value with the basic network (triangles) and with the cascaded network (circles). The training and testing sets are L_{NET2} and T_{37} (see Table 2)

Comparison with other methods on a protein basis

The predictive capability of neural networks on the never before seen patterns of some membrane proteins, representative or relevant functional classes, is compared with the results of two recent predictive methods (MSA and MEMSAT described in Persson and Argos 1994 and Jones et al. 1994, respectively). The performances obtained using MSA (drawn from Persson and Argos 1994), our method (CNET2/CUT2) and MEMSAT are compared in Table 10, in terms of the number and positions of the transmembrane segments predicted with each method. The structural models proposed in the literature for these proteins rely mainly on hydrophobicity plots and are still tentative (Holm et al. 1987; Miller 1991; Walker 1992). As the only exception, K^+-voltage ion channels have been remodelled on the basis of data obtained with recombinant DNA manipulation (Miller 1991). Their topology seems to be characterized by six α-helix transmembrane segments plus two unusual and probably non-helical membrane-spanning stretches, located between helix 5 and 6 and corresponding to the ion pore forming region. The neural predictor provides the closest prediction to the topography of the protein which is inserted in the membane so

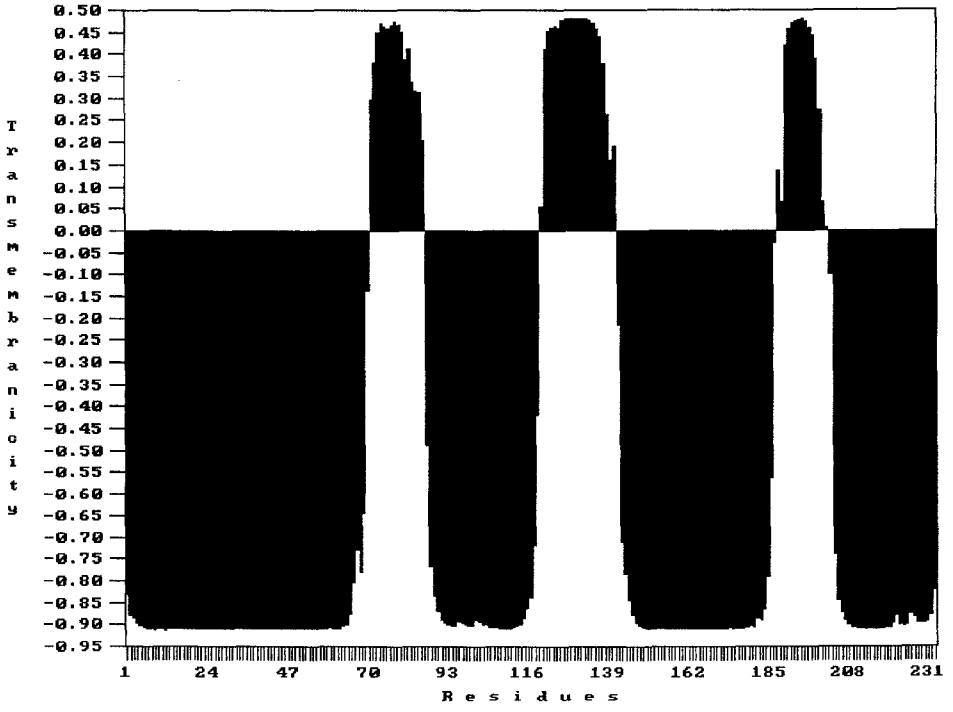

Fig. 5. Prediction of the transmembrane domains of the light-harvesting complex. The output (transmembranicity) is evaluated with CNET2 and plotted along the residue sequence (Kühlbrandt et al. 1994)

Table 10. Testing membrane proteins with different predictive methods

Protein		CNET2 (CUT2)[a]	MSA[b]	MEMSAT[c]
Voltage-gated	1:	184–202	182–205	183–201
K^+-channel	2:	246–263	242–262	243–264
(Cik3_Human)	3:	276–292	266–286	275–292
	4:	306–329	306–324	
	5:	331–363	341–367	346–367
	6:		377–395	
	7:	405–423	404–430	407–428
Cytochrome *c* oxidase	1:	15–39	12–40	15–37
polypeptide I	2:	54–86	59–87	67–87
(Cox1_Human)	3:	103–119	96–117	102–118
	4:	145–170	141–169	145–169
	5:	184–209	181–209	183–207
	6:	241–257	230–258	243–261
	7:	273–293	268–296	268–289
	8:	307–328	302–325	303–327
	9:	330–363	336–364	338–359
	10:	375–399	372–400	379–403
	11:	412–431	415–442	410–426
	12:	452–478	447–470	452–473
Cytochrome *c* oxidase	1:		9–29	
polypeptide III	2:	19–53	36–53	24–48
(Cox3_Human)	3:	83–103	80–103	81–101
	4:	130–147	131–150	131–147
	5:	158–177	155–182	160–176
	6:	197–226	193–221	196–220
	7:	243–260	234–258	237–254
NADH-ubiquinone	1:	1–27	1–29	7–23
oxidoreductase chain 1	2:	70–86	67–95	68–87
(Nu1m_Human)	3:	97–122	97–124	100–122
	4:	125–163	134–161	144–166
	5:	165–190	165–193	173–191
	6:			207–224
	7:	225–243	217–244	231–247
	8:	259–274	258–283	254–273
	9:	289–309	287–313	293–311
NADH-ubiquinone	1:	9–42	2–22	7–23
oxidoreductase chain 2	2:		27–47	30–46
(Nu2m_Human)	3:	60–82	61–81	59–76
	4:	92–106	88–108	
	5:	125–139	115–135	
	6:	144–167	142–162	151–170
	7:	176–189		178–194
	8:		187–212	
	9:	201–237		201–221
	10:	240–252	237–265	237–261
	11:	275–305	272–300	275–294
	12:	327–346	321–341	323–340
NADH-ubiquinone	1:	1–25	1–27	7–24
oxidoreductase chain 3	2:	54–81	50–78	55–79
(Nu3m_Human)	3:	83–106	82–106	86–104
NADH-ubiquinone	1:	1–18	3–24	8–24
oxidoreductase chain 5	2:	37–54	37–65	39–57
(Nu5m_Human)	3:	80–103	78–106	84–108
	4:	118–138	121–149	117–134
	5:	140–161		141–160
	6:	172–192	171–196	172–191
	7:	209–224	209–223	201–222
	8:	245–265	244–265	243–261
	9:	271–293	272–293	275–293
NADH-ubiquinone	10:	302–315	304–330	301–319
oxidoreductase chain 5	11:	318–344	335–359	326–347
(Nu5m_Human)	12:	366–383	364–389	373–389
	13:	404–430	401–427	406–422
	14:	454–468	454–477	450–469
	15:	483–505	481–509	480–502
	16:			517–533
	17:	540–560		
	18:	569–602	581–603	
Beta-glucosides-	1:	101–131	101–129	100–121
specific phospho-	2:	142–162	143–163	140–161
transferase enzyme II	3:	175–187	167–187	172–190
(Pt2b_Ecoli)	4:	207–234	216–236	206–230
	5:	244–263	248–268	245–269
	6:	265–290		
	7:	292–314	280–308	
	8:	316–341	319–345	324–346
	9:		357–377	
	10:	384–400	392–412	383–400
	11:	427–452	421–450	426–449
	12:	490–510		
	13:	524–534		

[a] CNET2 (CUT2) = Cascaded NET2 with CUT2 (this work)
[b] MSA = Method based on multiple sequence alignments (Persson and Argos 1994)
[c] MEMSAT = Method based on model recognition approach (Jones et al. 1994). Protein sequences are from the SWISS-PROT data bank

as to expose the N and C terminus on the inner side of the osmotic space, and miss as expected the putative non-helical region. As for the other proteins, including chain I and III of human cytochrome *c* oxidase (Holm et al. 1987) and chains 1, 2, 3 and 5 of human NADH-ubiquinone oxidoreductase (Walker 1992), the results indicate, with very few exceptions, a fairly good agreement between the numbers and locations of transmembrane segments predicted by the neural network and the other two statistical methods.

Discussion

In the present work the prediction of the transmembrane organization of membrane proteins is accomplished by using neural networks to determine the connections between the residue sequences and the transmembrane α-helical domains of the proteins of the training set. The rules of association are then used to make a blind prediction of the patterns of the proteins of the testing set. Probably the most striking result that we obtain is that a good level of prediction accuracy, as compared to expectations, is already obtained, notwithstanding the small size of the training set of the proteins known at atomic resolution. Apparently, the features pertaining to the transmembrane regions of

the proteins used are shared by the membrane proteins contained in the testing set.

The notion that topological determinants are common to most of the membrane proteins is at the basis of the predictive methods so far developed. Neural networks lend further support to this view and confirm that transmembrane domains are strictly connected to the residue sequence.

The advantage of using neural networks rather than other methods to cope with this predictive task extends beyond the high computational rates provided by massive parallelism. These devices, being composed of many nonlinear computational elements operating in parallel, are capable of learning a general solution to a problem from a set of specific examples. During the supervised training phase, the connection weights are adapted iteratively to improve the performance obtained on the proteins of the training set. In doing so, the specific patterns pertaining to transmembrane domains of membrane proteins are memorized in the connections of the networks while the sliding input window scans the protein sequences. In contrast to classical statistical methods based on single-residue information, this procedure automatically accounts for the relative positions of the residues within the sliding window, as is shown in the Hinton diagram of Fig. 3. This is reminiscent of the directional information based on information theory and underlying the propensities for different secondary structures in globular proteins (Gibrat et al. 1987; Garnier and Robson 1989). A further characteristic of this approach is that the learning phase amounts to the automatic computation of a decision function; this reduces to a minimum the *a posteriori* tailoring procedure used to optimize several of the methods predicting membrane proteins.

When training is performed on sets of small size, a free variable which must be carefully controlled is the number of training cycles, namely the number of times the training set is presented to the network. With small training sets, after a limited number of training cycles, the network already memorizes the specific patterns of the target proteins of the training set, as indicated by the large values (approaching 1) of the accuracy indices obtained by testing the training sets (Fariselli et al. 1993). However, upon attaining this regime, the generalization capability of the network on the never-seen-before proteins of the testing set degrades. It is therefore necessary to find the optimal number of training cycles in order to ensure a satisfactory prediction efficiency. This is heuristically determined with a search in the parameter space of the networks. When generalizing on bacteriorhodopsin, the maximum performance of the network is reached after 30 cycles of training, with a window size of 17 residues and a learning rate of 0.01. The optimal window size turns out to be comparable with the typical length of transmembrane segments and in the range used before to optimize average hydrophobicity values (Kyte and Doolittle 1982; Engelman et al. 1986).

The main characteristics of the neural network based predictor are discussed in the following.

The predictor is highly specific for α-helix domains of membrane proteins. This is indicated by the low rate of false positive assignments on testing globular proteins. By applying the filtering procedure (directly calibrated on the testing performance of the predictor on the proteins of the training set) the mispredicted cases out of 26 826 residues of globular proteins belonging to different structural classes are reduced within the range of 1.6–3.5%. This error rate ranks lower than that obtained with the most recent predictive methods (Edelman 1993; Jones et al. 1994) and much lower than that obtained with other statistical or empirical methods (Edelman 1993). Moreover the predictor correctly does not predict the β-strand rich transmembrane segments of porin.

When tested on membrane proteins representative of different functional classes with no homology with those of the training set (L_{NET1}), the predictive accuracy is higher by two percentage points than that obtained with a predictor based on quadratic minimization of the window and decision functions, and tuned on the same training set of membrane proteins (Edelman 1993). It must, however, be considered that in the absence of crystallographic data, most of the models so far described for membrane proteins are still debated. As a consequence the computed statistical parameters are only indicative. The predictive accuracy can be as high as 90% of total predictions and in fairly good agreement with the expected locations of the transmembrane segments, as indicated by the high score of the correlation coefficients and overlapping indexes.

The average length of the predicted transmembrane segments is quite close to the typical value for a transmembrane α-helix, although transmembrane stretches of different sizes can be recognized by the network.

A further parameter of interest that can be directly estimated from the neural network's outputs is the reliability of the prediction. This value is 100% for 60% of the predicted 18 242 residues of the testing set, indicating a quite high score for the overall performance.

Despite confident predictions of some adrenergic and acetylcholine receptors (Peralta et al. 1987; Strosberg 1991), the predictor in its present version fails to recognize the seventh helix of rhodopsin and other related proteins of the G-coupled receptors. The signals pertaining to this segment, although present, are weak and rejected also by the less stringent filtering procedure (CUT2). This indicates that this pattern, if we accept that it corresponds to a membrane segment (Dohlman et al. 1991), is dissimilar to all those used to train the network. The inclusion of some amphipathic α-helices in the training set is not sufficient to change this result. The prediction of the seventh helix of the G-coupled receptors is improved only by including in the training set the more relevant proteins of this family (data not shown). Interestingly the seventh α-helix of the G-coupled receptors is also weakly predicted or missed by the predictors based on model recognition (Jones et al. 1994) and on evolutionary information (Persson and Argos 1994), respectively. The fact that these methods are based on totally different approaches, as compared to neural networks, confirms the notion that the pattern of this α-helix is rather unusual as compared to typical transmembrane domains.

When experimental data are not yet available to support the possible models of membrane proteins, the con-

vergence of different predictors to similar models may be particularly significant. An example is given when the predictive performance of neural networks is compared on a protein basis with that of MSA (Persson and Argos 1994) and of MEMSAT (Jones et al. 1994). These three predictors lead to quite similar results (Table 10).

In conclusion, our analysis provides evidence that the predictive accuracy and the correct location of transmembrane segments along the protein sequence are dependent on the extent of the directional information automatically extracted by the network from the proteins of the training set. In a previous work (Fariselli et al. 1993), we have shown that traits of secondary structure are common to globular and membrane proteins. The predictive accuracy of secondary structures of crystallized membrane proteins is about 70% when the training set comprises all-α globular proteins (Compiani et al. 1991), indicating that α-helical motifs can be extrapolated from globular to membrane proteins. A more constrained task is however addressed in the present work, since the network is trained to recognize only transmembrane α-helices. The increase of about twenty percentage points registered in the prediction accuracy for α-helices embedded in the bilayer can be traced back to the connections between the protein sequence and the transmembrane domains that are fully exploited in the present investigation.

Acknowledgements. We thank Dr. B. Rost (EMBL, Heidelberg, Germany) for discussions. We are also indebted to Dr. D. T. Jones for sending us the program MEMSAT and sequences of membrane proteins. This work was partially supported by the Consiglio Nazionale delle Ricerche.

References

Argos P, Rao JKM, Hargrave PA (1982) Structural prediction of membrane-bound proteins. Eur J Biochem 128:565–575

Bairoch A, Boeckmann B (1992) The SWISS-PROT protein sequence data bank. Nucl Acids Res 20:2019–2022

Bernstein FC, Koetzle TF, Williams GJB, Meyer EF, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M (1977) The protein data bank: a computer-based archival file for macromolecular structures. J Mol Biol 112:535–542

Black SD, Coon MJ (1982) Structural features of liver microsomal NADPH-cytochrome P-450 reductase. J Biol Chem 257:5929–5938

Bowie JU, Luthy R, Eisenberg DA (1991) Method to identify protein sequences that fold into a known three-dimensional structure. Science 235:164–170

Brandl CJ, Green NM, Korczak B, MacLennan DH (1986) Two Ca^{2+} ATPase genes: homologies and mechanistic implication of deduced amino acid sequences. Cell 44:597–607

Brunisholz RA, Zuber H, Valentine J, Lindsay JG, Woolley KJ, Cogdell RJ (1986) The membrane location of the B890-complex from *Rhodospirillum rubrum* and the effect of carotenoid on the conformation of its two apoproteins exposed at the cytoplasmic surface. Biochim Biophys Acta 849:925–303

Burgi R, Suter F, Zuber H (1987) Arrangement of the light-harvesting chlorophyll a/b protein complex in the thylakoid membrane. Biochim Biophys Acta 890:346–351

Chothia C (1992) One thousand families for the molecular biologist. Nature 357:543–544

Chothia C, Finkelstein AV (1990) The classification and origins of protein folding patterns. Annu Rev Biochem 59:1007–1039

Compiani M, Fariselli P, Casadio R (1991) Neural networks extracting general features of protein secondary structures. In: Caianiello ER (ed) Parallel Architectures and Neural Networks. World Scientific, Singapore, pp 227–237

Compiani M, Fariselli P, Casadio R (1992) The statistical behaviour of perceptrons. In: Caianiello ER (ed) WIRN Vietri-92. World Scientific, Singapore, pp 111–117

Cornette JL, Cease KB, Margalit H, Spouge JL, Berzofsky JA, De Lisi C (1987) Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. J Mol Biol 195:659–685

Degli Esposti M, Crimi M, Venturoli G (1990) A critical evaluation of the hydropathy profile of membrane proteins. Eur J Biochem 190:207–219

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

Dohlman HG, Thorner J Caron, Lefkowitz M (1991) Model systems for the study of the seven-transmembrane segment receptors. Annu Rev Biochem 60:653–688

Edelman J (1993) Quadratic minimization of predictors for protein secondary structure: application to transmembrane α-helixes. J Mol Biol 232:165–191

Edelman J, White SH (1989) Linear optimization of predictors for secondary structure: application to transbilayer segments of membrane proteins. J Mol Biol 210:195–209

Eisenberg D, Schwarz E, Komaromy M, Wall R (1984) Analysis of membrane and surface protein sequences with the hydrophobic moment plot. J Mol Biol 179:125–142

Engelman DM, Steitz TA, Goldman A (1986) Identifying nonpolar transbilayer helixes in amino acid sequences of membrane proteins. Annu Rev Biophys Biophys Chem 15:321–353

Fariselli P, Compiani M, Casadio R (1993) Predicting secondary structures of membrane proteins with neural networks. Eur Biophys J 22:41–51

Fasman GD (1989) The development of the prediction of protein structure. In: Fasman GD (ed) Prediction of Protein Structure and the Principles of Protein Conformation. Plenum Press, New York, pp 193–316

Fasman GD, Gilbert WA (1990) The prediction of transmembrane protein sequences and their conformation: an evaluation. TIBS 15:89–92

Feher G, Allen JP, Okamura MY, Rees DC (1989) Structure and function of bacterial photosynthetic reaction centers. Nature 339: 111–116

Foster DL, Boublik M, Kaback HR (1983) Structure of the *lac* carriers protein of *Escherichia coli.* J Biol Chem 258:31–34

Fox RO, Richards FM (1982) A voltage gated ion channel model inferred from the crystal structure of alamethicin at 1.5 Å resolution. Nature 300:325–330

Fujii-Kuriyama Y, Mizukami Y, Kawajiri K, Sogawa K, Muramatsu M (1982) Primary structure of a cytochrome P-450: coding nucleotide sequence of phenobarbital-inducible cytochrome P-450 messenger RNA from rat liver. Proc Natl Acad Sci, USA 79:2793–2797

Garnier J, Robson B (1989) The GOR method for predicting secondary structures in proteins. In: Fasman GD (ed) Prediction of Protein Structure and the Principles of Protein Conformation. Plenum Press, New York, pp 417–465

Garnier J, Levin JM (1991) The protein code: what is the present status? CABIOS 7:133–142

Gibrat JF, Garnier J, Robson B (1987) Further developments of protein secondary structure prediction using information theory. J Mol Biol 198:425–443

Gimlich RL, Kumar NM, Gilula NB (1988) Sequence and developmental expression of mRNA coding for a gap junction protein in *Xenopus*. J Cell Biol 107:1065–1073

Grenningloh G, Rienitz A, Schmitt B, Methfessel C, Zensen M, Beyreuter K, Gundelfingr ED, Betz H (1987) The strychnine-bind-

ing subunit of the glycine receptor shows homology with nicotinic acetylcholine receptors. Nature 328:215–220
Henderson R, Baldwin JM, Ceska TA, Zemlin F, Beckmann E, Downing KH (1990) Model for the structure of bacteriorhodopsin based on high resolution electron cryo-microscopy. J Mol Biol 213:899–929
Hirst JD, Sternberg JE (1992) Prediction of structural and functional features of proteins and nucleic acid sequences by artificial neural networks. Biochemistry 31:7211–7218
Holley HL, Karplus M (1989) Protein secondary structure prediction with a neural network. Proc Natl Acad Sci, USA 85: 152–156
Holm L, Saraste M, Wilkstrom M (1987) Structural models of the redox centres in cytochrome oxidase. EMBO J 6:2819–2823
Jennings ML (1989) Topography of membrane proteins. Annu Rev Biochem 58:999–1027
Jones DT, Taylor WR, Thornton JM (1994) A model recognition approach to the prediction of all-helical membrane protein structure and topology. Biochemistry 33:3038–3049
Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2673
Kayano T, Noda M, Flockerzi V, Takahashi H, Numa S (1988) Primary structure of rat brain sodium channel III deduced from the cDNA sequence. FEBS Letters 228:187–194
Khorana HG (1992) Rhodopsin, photoreceptor of the rod cell. J Biol Chem 267:1–4
Klein P, Kanehisa M, De Lisi C (1985) The detection and classification of membrane-spanning proteins. Biochim Biophys Acta 815:468–476
Kneller DG, Cohen FE, Langridge R (1990) Improvements in protein secondary structure prediction by an enhanced neural network. J Mol Biol 214:171–182
Kopito RR, Lodish HF (1985) Primary structure and transmembrane orientation of the murine anion exchange protein. Nature 316:234–238
Kyte J, Doolittle RF (1982) A simple method for displaying the hydrophobic character of a protein. J Mol Biol 157:103–132
Kühlbrandt W, Wang DN, Fujiyoshi Y (1994) Atomic model of plant light-harvesting complex by electron crystallography. Nature 367:614–621
Kuhn LA, Leigh JS (1985) A statistical technique for predicting membrane protein structure. Biochim Biophys Acta 828: 351–361
Lipman DJ, Pearson WL (1985) Rapid and sensitive protein similarity searches. Science 227:1435–1441
Lippmann RP (1987) An introduction to computing with neural nets. IEEE ASSP Magazine 4:4–22
Manoil C, Beckwith J (1986) A genetic approach to analyzing membrane protein topology. Science 233:1403–1408
Matthews BW (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 405:442–451
McGovern K, Ehrmann M, Beckwith J (1991) Decoding signals for membrane proteins using alkaline phosphatase fusions. EMBO J 10:2773–2782
Mellor IR, Thomas DH, Sansom MSP (1988) Properties of ion channels formed by *Staphylococcus aureus* δ-toxin. Biochim Biophys Acta 942:280–294
Michel H (1983) Crystallization of membrane proteins. Trends Biochem Sci 8:56–59
Michel H, Weyer KA, Gruenberg H, Dunger I, Osterhelt D, Lottspeich F (1986) The light and medium subunits of the photosynthetic reaction center from *Rhodopseudomonas viridis:* isolation of the genes, nucleotide and amino acid sequence. EMBO J 5:1149–1158
Miller C (1991) 1990: annus mirabilis of potassium channels. Science 252:1092–1096
Moore KE, Miura SA (1987) Small hydrophobic domain anchors leader peptidase to the cytoplasmic membrane of *Escherichia coli*. J Biol Chem 262:8806–8813
Müller D, Reinhardt J (1990) Neural networks. Springer, Berlin Heidelberg New York
Noda M, Takahashi H, Tanabe T, Toyosato M, Kikyotani S, Furutani Y, Hirose T, Takashima H, Inayama S, Miyata T, Numa S (1983a) Structural homology of *Torpedo californica* acetylcholine receptor subunits. Nature 302:528–532
Noda M, Takahashi H, Tanabe T, Toyosato M, Kikyotani S, Hirose T, Asai M, Takashima H, Inayama S, Miyata T, Numa S (1983b) Primary structure of β- and δ-subunit precursors of *Torpedo californica* deduced from cDNA sequences. Nature 301:251–255
Ozols J, Carr SA, Strittmatter P (1984) Identification of the NH_2-terminal blocking group of NADH-cytochrome b_5 reductase as myristic acid and the complete amino acid sequence of the membrane binding domain. J Biol Chem 259:13349–13354
Peralta EG, Ashkenazi A, Winslow JW, Smith DH, Ramachandran J, Capon DJ (1987) Distinct primary structures, ligand-binding properties and tissue specific expression of four human muscarinic acetylcholine receptors. EMBO J 6:3923–3929
Persson B, Argos P (1994) Prediction of transmembrane segments in proteins utilising multiple sequence alignments. J Mol Biol 237:182–192
Popot JL, De Vitry C (1990) On the microassembly of integral membrane proteins. Annu Rev Biophys Biophys Chem 19:369–403
Popot JL, Dinh DP, Dautigny A (1991) Major myelin proteolipid: the 4-α-helix topology. J Membr Biol 120:233–246
Presnell SR, Cohen FE (1993) Artificial neural networks for pattern recognition in biochemical sequences. Annu 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 202:865–884
Rao JKM, Argos P (1986) A conformational preference parameter to predict helices in integral membrane proteins. Biochim Biophys Acta 869:197–214
Ross AH, Radhakrishnan R, Robson RJ, Khorana HG (1982) The transmembrane domain of glycophorin A as studied by crosslinking using photoactivable phospholipids. J Biol Chem 257: 4152–4161
Rost B, Sander C (1993) Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol 232:584–599
Rost B, Sander C, Schneider R (1994) Redefining the goals of protein secondary structure prediction. J Mol Biol 235:13–26
Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representation by back-propagation errors. Nature 323:533–536
Sayre RT, Andersson B, Bogorad L (1986) The topology of a membrane protein: the orientation of the 32 kD Qb-binding chloroplast thylakoid membrane proteins. Cell 47:601–608
Schofield PR, Darlison MG, Fujita N, Burt DR, Stephenson FA (1987) Sequence and functional expression of the $GABA_A$ receptor shows a ligand-gated receptor super-family. Nature 328: 221–227
Schulz GE (1988) A critical evaluation of methods for prediction of protein secondary structures. Annu Rev Biophys Chem 17:1–21
Schulz GE, Schirmer RH (1979) Principles of protein structure. Springer, Berlin Heidelberg New York
Shull GE, Lane LK, Lingrel JB (1986) Amino acid sequence of the β-subunit of the $(Na^+ + K^+)$ATPase deduced from cDNA. Nature 321:429–431
Stader J, Matsumura P, Vacante D, Dean DE, Macnab RM (1986) Nucleotide sequence of the *Escherichia coli motB* gene and site-limited incorporation of its product into the cytoplasmic membrane. J Bacteriol 166:244–252
Strosberg AD (1991) Structure/function relationship of proteins belonging to the family of receptors coupled to GTP-binding proteins. Eur J Biochem 29:11009–11023
Stumher W, Ruppersberg JP, Schroter KH, Sakmann B, Stoker M, Giese KP, Perschke A, Baumann A, Pongs O (1989) Molecular basis of functional diversity of voltage-gated potassium channels in mammalian brain. EMBO J 8:3235–3244

Von Heijne G (1988) Trascending the impenetrable: how proteins come to term with membrane. Biochim Biophys Acta 947: 307–333

Von Heijne G (1992) Hydrophobicity analysis and the positive-inside rule. J Mol Biol 225:487–494

Takagaki Y, Radhakrishnan R, Gupta CM, Khorana HG (1983) The membrane-embedded segment of cytochrome b_5 as studied by cross-linking with photoactivable phospholipids. I. The transferable form. J Biol Chem 258:9128–9135

Tanabe T, Takeshima H, Mikami A, Flockerzi V, Takahashi H (1987) Primary structure of the receptor for calcium channel blockers from skeletal muscle. Nature 328:313–318

Terwillinger TC, Weissman L, Eisenberg D (1982) The structure of melittin in the form I crystals and its implication for melittin's lytic and surface activities. Biophys J 27:353–361

Traxler B, Boyd D, Beckwith J (1993) The topological analysis of integral membrane proteins. J Membr Biol 132:1–11

Walker JE (1992) The NADH:ubiquinone oxidoreductase (complex I) of respiratory chains. Q Rev Biophys 25:253–324

Weiss MS, Kreush A, Shiltz E, Nestel U, Welte W, Weckesser J, Schulz GE (1991) The structure of porin from *Rhodobacter capsulatus* at 1.8 Å resolution. FEBS Lett 280:379–382