Protein Science (1995), *4*:521–533. Cambridge University Press. Printed in the USA.

Transmembrane helices predicted at 95% accuracy

BURKHARD ROST,[1] RITA CASADIO,[2] PIERO FARISELLI,[2] AND CHRIS SANDER[1]

[1] Protein Design Group, EMBL Heidelberg, 69 012 Heidelberg, Germany
[2] Laboratory of Biophysics, Department of Biology, University of Bologna, 40 126 Bologna, Italy

(RECEIVED October 31, 1994; ACCEPTED December 29, 1994)

Abstract

We describe a neural network system that predicts the locations of transmembrane helices in integral membrane proteins. By using evolutionary information as input to the network system, the method significantly improved on a previously published neural network prediction method that had been based on single sequence information. The input data were derived from multiple alignments for each position in a window of 13 adjacent residues: amino acid frequency, conservation weights, number of insertions and deletions, and position of the window with respect to the ends of the protein chain. Additional input was the amino acid composition and length of the whole protein. A rigorous cross-validation test on 69 proteins with experimentally determined locations of transmembrane segments yielded an overall two-state per-residue accuracy of 95%. About 94% of all segments were predicted correctly. When applied to known globular proteins as a negative control, the network system incorrectly predicted fewer than 5% of globular proteins as having transmembrane helices. The method was applied to all 269 open reading frames from the complete yeast VIII chromosome. For 59 of these, at least two transmembrane helices were predicted. Thus, the prediction is that about one-fourth of all proteins from yeast VIII contain one transmembrane helix, and some 20%, more than one.

Keywords: evolutionary information; integral membrane proteins; multiple alignments; neural networks; protein structure prediction; secondary structure; yeast VIII chromosome

Given the rapid advance of large-scale gene-sequencing projects (Oliver et al., 1992; Johnston et al., 1994), most protein sequences of key organisms will be known in about 5 years' time. Experimental structure determination is becoming more of a routine (Lattman, 1994); and the number of proteins with known sequence for which the three-dimensional (3D) structure can be predicted rather accurately by homology modeling is constantly increasing (today more than 25% of all sequences in the SWISS-PROT sequence data base [Bairoch & Boeckmann, 1994] can be modeled with reasonable accuracy by homology [Sander & Schneider, 1994]). Even in such an optimistic scenario, experimental knowledge about membrane proteins is likely to be sparse. However, membrane proteins represent a very important class of protein structures. To what extent can structural aspects for membrane proteins be predicted from sequence information?

Two types of membrane proteins. So far, the 3D structures of two types of membrane proteins have been determined. The first type are helical proteins: photosynthetic reaction center (Deisenhofer et al., 1985), bacteriorhodopsin (Henderson et al., 1990), and the light harvesting complex II (Wang et al., 1993; Kühlbrandt et al., 1994); these proteins consist of typically apolar helices of some 20 residues that traverse the membrane perpendicular to its surface (Fig. 1). The second type is represented by the structure of porin (Weiss & Schulz, 1992; Cowan & Rosenbusch, 1994), a 16-stranded β-barrel.

Membrane proteins easier to predict than globular ones. Typical methods for the prediction of transmembrane segments focus on helical transmembrane (HTM) proteins (von Heijne, 1981, 1986; Argos et al., 1982; Eisenberg et al., 1984a; Engelman et al., 1986; von Heijne & Gavel, 1988). It is commonly believed that the prediction of structure is simpler for membrane proteins than for globular ones as the lipid bilayer imposes strong constraints on the degrees of freedom of structure (Taylor et al., 1994).

Prediction of transmembrane segments. Methods for prediction of transmembrane helices are usually based on (1) hydrophobicity analyses (Argos et al., 1982; Kyte & Doolittle, 1982; Engelman et al., 1986; Cornette et al., 1987; Degli Esposti et al., 1990); (2) the preponderance of positively charged residues on the cytoplasmic side of the transmembrane segment (interior), established as the "positive inside rule" (von Heijne, 1981, 1986, 1991, 1992; von Heijne & Gavel, 1988; Sipos & von Heijne, 1993); or (3) statistical procedures that perform significantly bet-

Reprint requests to: Burkhard Rost, Protein Design Group, EMBL Heidelberg, 69 012 Heidelberg, Germany; e-mail: rost@embl-heidelberg.de.

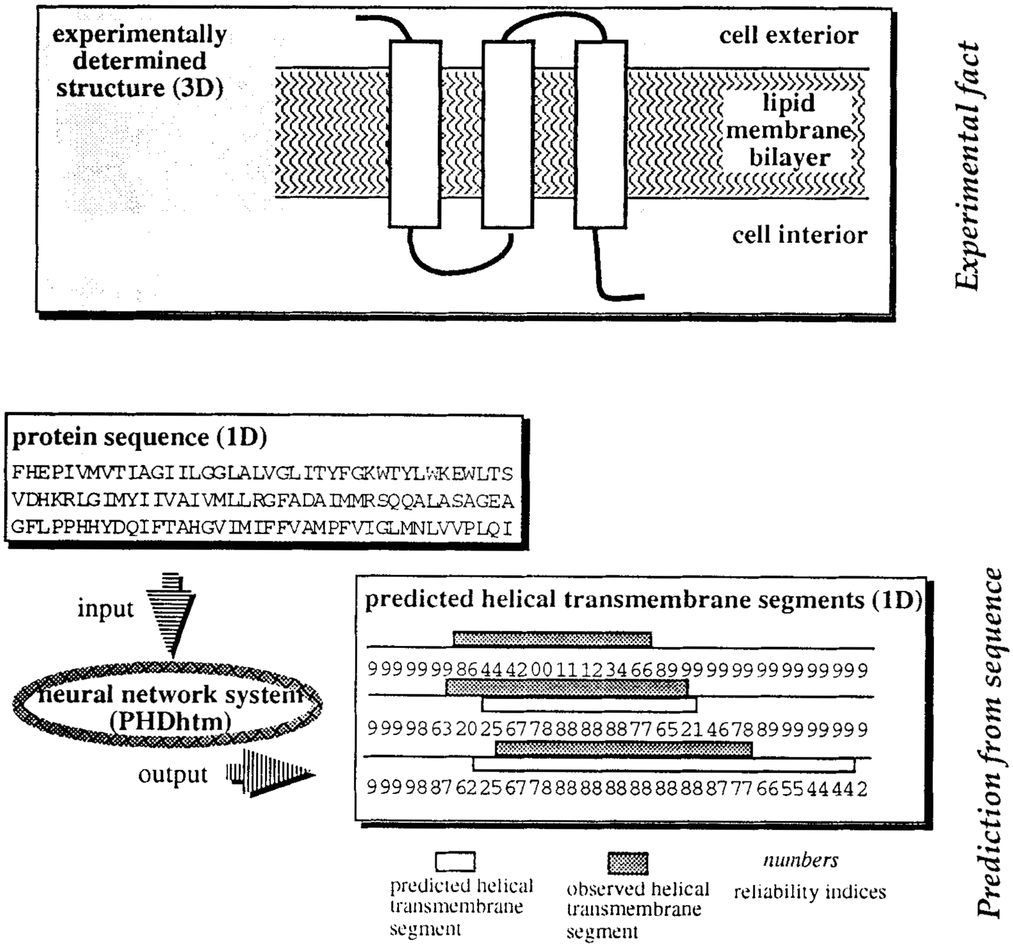

experimentally determined structure (3D)
cell exterior
lipid membrane bilayer
cell interior
Experimental fact
protein sequence (1D)
FHEPIVMVTIAGIILGGLALVGLITYFGKWTYLWKEWLTS
VDHKRLGIMYIIVAIVMLLRGFADAIMMRSQQALASAGEA
GFLPPHHYDQIFTAHGVIMIFFVAMPFVIGLMNLVVPLQI
input
neural network system (PHDhtm)
output
predicted helical transmembrane segments (1D)
9999999864442001112346689999999999999999
9999863202567788888877652146788999999999
9999887622567788888888888887776655444442
predicted helical transmembrane segment
observed helical transmembrane segment
numbers
reliability indices
Prediction from sequence

Fig. 1. Prediction of the location of transmembrane helices. In one class of membrane proteins, typically apolar helical segments are embedded in the lipid bilayer oriented perpendicular to the surface of the membrane. Helical segments can be regarded as more or less rigid cylinders. Thus, the 3D structure of the membrane spanning protein region can be determined by: the location of segments with respect to sequence; the orientation of helical axes; the inclination of helical axes with respect to lipid bilayer; and the phase of helices with respect to each other (orientation of helical wheel). Here, we simplify extremely by projecting 3D structure onto a 1D string describing which residues of the protein are part of a transmembrane helices. Input to the prediction tool (neural network system) is a protein sequence (in general a sequence alignment), output is a prediction of the location of transmembrane segments. The example shown (sequence of cytochrome O ubiquinol oxidase subunit I, cyob_eco in SWISS-PROT; Bairoch & Boeckmann, 1994) contained one of the few segments that were underpredicted (missed). The numbers give the reliability of the prediction for each residue on a scale of 0–9 (Fig. 2). Nontransmembrane regions, when predicted correctly, usually reached the highest reliability (9). Thus, the unusually low reliability values for the underpredicted segment might have enabled the expert user to improve the automatic prediction by interpreting this region as nonloop.

ter when combined with multiple alignments (Persson & Argos, 1994). In general, prediction of transmembrane segments is relatively straightforward. But, can detailed aspects of 3D structure be predicted from sequence for HTM proteins?

Prediction of 3D structure for HTM proteins. Cytoplasmic and extracellular regions have different amino acid compositions (von Heijne & Gavel, 1988; Nakashima & Nishikawa, 1992). This difference allows for a successful prediction of not only the location of helices but, as well, of their orientation with respect to the cell (pointing inside or outside the cell) (Landolt-Marticorena et al., 1992; Sipos & von Heijne, 1993; Jones et al., 1994). Going further, Taylor and colleagues enumerate all possible models for packing seven-helix transmembrane proteins and select the "better models" (Taylor et al., 1994). The selection criterion for "better models" is the crucial point of the method. The authors report that the native conformation is found in "most cases" tested. However, the N- and C-terminal ends of the transmembrane helices have to be predicted very accurately for a successful automatic prediction of 3D structure from sequence (Taylor et al., 1994). Can the accuracy of predicting not just the location of transmembrane helices but, as well, of the N- and C-terminal ends be improved?

Better prediction of transmembrane helix location. Prediction accuracy has recently been improved significantly (Sipos & von Heijne, 1993; Jones et al., 1994; Persson & Argos, 1994). A system of neural networks using single sequences as input (Fariselli et al., 1993; R. Casadio, P. Fariselli, C. Taroni, & M. Compiani, submitted for publication) appears to be slightly inferior to these methods. However, using information from multiple sequence alignments as input, neural networks have been shown to yield the most accurate prediction of secondary structure for globular proteins (Rost & Sander, 1993a, 1993c, 1994a). Here, we used a similar system of neural networks to predict transmembrane helices based on evolutionary information (Figs. 1, 2). The goal was to predict the location of transmembrane helices (defined as helix caps given in SWISS-PROT [Bairoch & Boeckmann, 1994]) more accurately than alternative methods (Sipos & von Heijne, 1993; Jones et al., 1994; Persson & Argos, 1994; R. Casadio et al., submitted). The neural network system was tested in fivefold cross-validation on 69 proteins with experimentally well-determined transmembrane helices (Materials and methods). Network input was the information derived for successive windows of 13 adjacent residues from a multiple sequence alignment (Fig. 3). Output were two units, one for each state of the central residue (in membrane helix/not in membrane helix; Fig. 2).

Results and discussion

Evolutionary information improves prediction accuracy significantly

Better prediction in terms of per-residue and segment-based scores. Compared to a simple neural network, the per-residue accuracy of the full three-level system using explicitly various aspects of evolutionary information increased by some five percentage points (Table 1). The improvement in prediction accuracy was even more significant in terms of segment-based scores: from some 75% correctly predicted segments to 94%.

Reliability index of practical use to refine prediction accuracy. For some 70% of all proteins, 100% of all segments were predicted correctly (data not shown). The reliability of the prediction (reliability index defined in Fig. 4) can help to estimate whether or not a protein is likely to belong to the majority of proteins for which all segments are predicted correctly (Fig. 4). Furthermore, the reliability index was used to control the filtering procedure (Fig. 5).

Performance similar to that of the best alternative methods

Recently, two groups reported significant improvements in predicting transmembrane helices. Jones et al. (1994) use a new method with five output states (HTM-inside/middle/outside and not-HTM inside/outside, where inside/outside refers to inside/outside the cell). Persson and Argos (1994) use four output states (HTM-begin/middle/end and not-HTM) plus multiple alignment information. The system described here resulted in an accuracy in predicting the transmembrane helices similar to these two methods although we used only two output states. An exact comparison of the performance accuracy is made difficult because for both methods neither are per-residue scores published nor are the segment measures used defined (see footnotes to Table 1). Surprisingly, the errors made by the network system are often different from those made by the two statistical methods (Table 2 in comparison to Jones et al., 1994; Persson & Argos, 1994).

High reliability in discriminating between proteins with and without transmembrane helices

Does the prediction method distinguish transmembrane from nontransmembrane proteins? Two questions are of interest. First, did the network system correctly predict all transmembrane proteins used for the cross-validation analysis as transmembrane proteins? And second, were some globular proteins falsely predicted to contain transmembrane segments?

Transmembrane proteins correctly identified. Both the network system using single sequences as input and the network using only profiles identified all but two proteins in the test set as transmembrane proteins: melittin (2mlt) and immunoglobulin G-binding protein precursor (iggb_strsp). Melittin is a special case because the DSSP (Kabsch & Sander, 1983) assignment of secondary structure splits the long helix of the 26-residue molecule into two that were so short that the filtering procedure would miss this protein even on the basis of the known 3D structure. The ultimate network system PHDhtm missed only melittin; all other membrane proteins were correctly identified.

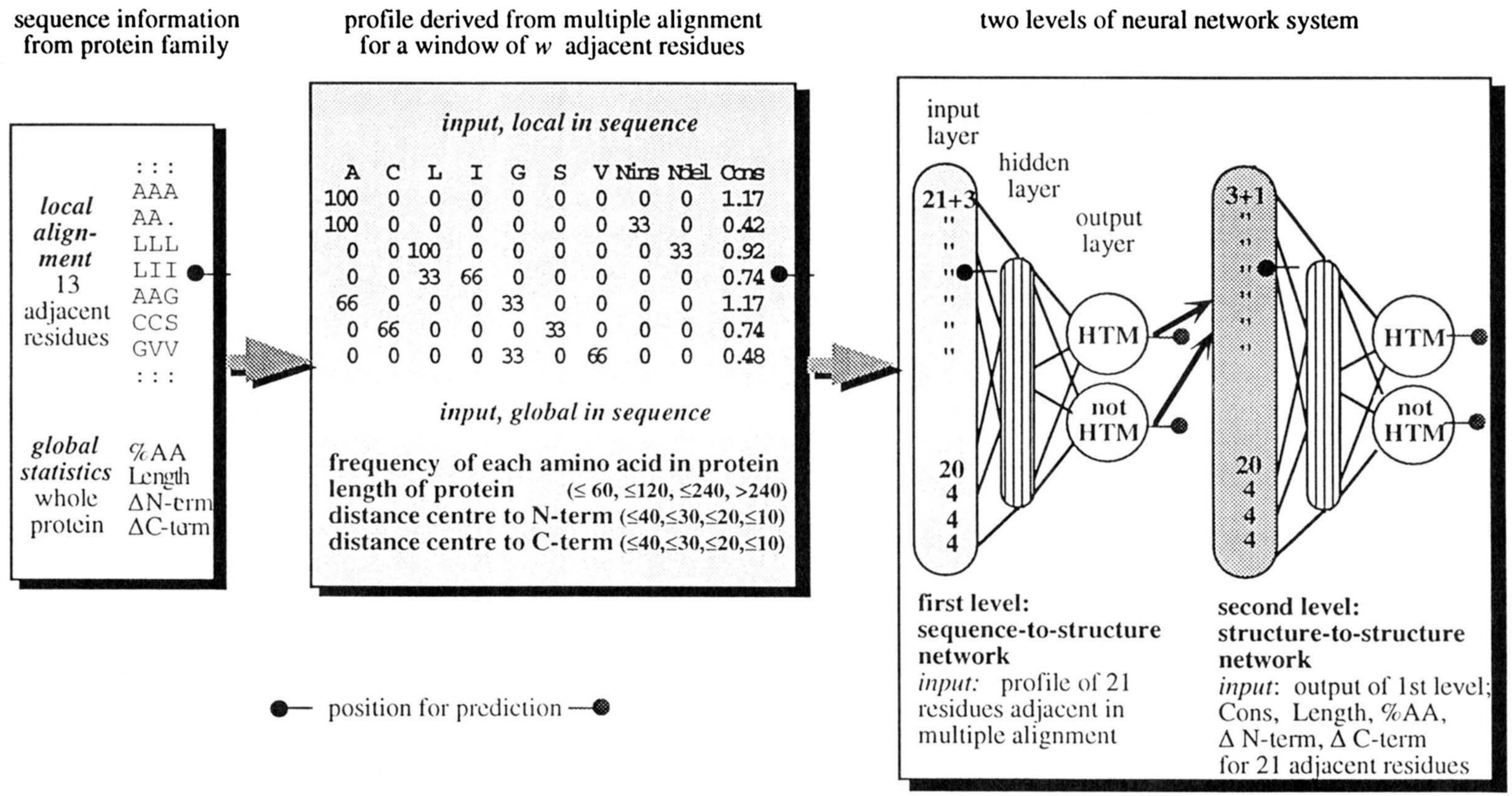
Fewer than 5% false positives. To test whether globular proteins were falsely predicted to contain transmembrane helices, we chose a set of 278 unique globular proteins. (No network predicted a transmembrane helix in the β-barrel porin.) PHDhtm mispredicted fewer than 5% of the globular proteins (Table 3). False positives were often globular water-soluble proteins with highly hydrophobic β-strands in the core. An exception was the only globular protein predicted to contain more than three segments: photosynthetic reaction center (4rcr) for which 11 segments with an average length of 21 residues were predicted as transmembrane helices (mandelate race mace [2mnr] was predicted with three long helices). The network using only profiles as input predicted transmembrane helices for less than 2% of the globular proteins.

Multilevel system improves significantly over simple neural network

Alignment information improves performance. The most significant improvement in prediction accuracy (compared to a simpler neural network prediction) stemmed from including the information contained in multiple alignments. Roughly one half of the improvement attributed to simply using residue substitution frequencies (Table 4), and one half to using additionally more details contained in the alignments (conservation weight, number of insertions and deletions) and information about the whole protein (Table 4).

Balanced versus unbalanced training. The balanced training procedure (equally often presenting residues in transmembrane and residues not in transmembrane segments; Materials and methods) tended to overpredict transmembrane helices, whereas an unbalanced training procedure (presentation of examples according to the distribution in the training set; Materials and methods) tended to underpredict transmembrane segments.

Jury decision finds a compromise between balanced and unbalanced training. Both balanced and unbalanced training had advantages and disadvantages. Which of the two methods should be used for prediction? A reasonable compromise (effectively between over- and underprediction) was found by the



A	C	L	I	G	S	V	Nins	Ndel	Cons
100	0	0	0	0	0	0	0	0	1.17
100	0	0	0	0	0	0	33	0	0.42
0	0	100	0	0	0	0	0	33	0.92
0	0	33	66	0	0	0	0	0	0.74
66	0	0	0	33	0	0	0	0	1.17
0	66	0	0	0	33	0	0	0	0.74
0	0	0	0	33	0	66	0	0	0.48

frequency of each amino acid in protein
length of protein (≤ 60, ≤120, ≤240, >240)
distance centre to N-term (≤40,≤30,≤20,≤10)
distance centre to C-term (≤40,≤30,≤20,≤10)

first level: sequence-to-structure network
input: profile of 21 residues adjacent in multiple alignment

second level: structure-to-structure network
input: output of 1st level; Cons, Length, %AA, Δ N-term, Δ C-term for 21 adjacent residues

Fig. 2. Two-level system of neural networks for HTM prediction. For each position in the alignment, the amino acid frequencies were compiled, the numbers of insertions and deletions counted, and a conservation weight computed (defined in Rost & Sander, 1993b). Furthermore, "global information" (beyond the window of 13 adjacent residues) about the search sequence was compiled: amino acid composition, length, and the position of the current window with respect to the N- and C-terminal end of the protein. All this information was fed into the neural network input for $w = 13$ adjacent residues (shown $w = 7$). The input layer was fully connected to a layer with three hidden units, and from there to the two output units coding for the central residues in the window (here "LII") to be in an HTM or not. The output of the first level was fed into a second level of structure-to-structure network, which additionally used the global information and the conservation weight as input. For this network, 15 hidden units were used. The two output units code again for the secondary structure state of the central residues (here "LII"). For first-level input units, local information is coded by $w \times (21 + 3)$ units, 20 for each amino acid, 1 for a spacer (for allowing windows to extend beyond protein ends, such that the first and last $w - 1$ residues in a protein can be used as central residue), and 3 for conservation weights, numbers of insertions, and numbers of deletions. Global information is coded by 32 additional units; 20 for the frequency of each amino acid in the protein, 4 for the length of the protein, and 4 for the distance of the central residue to the N- and 4 for the distance to the C-term of the protein. For second-level input units, the local information is coded by $w \times (3 + 1)$ units, two for each output unit of the first level (HTM, not HTM), one for a spacer, and one for the conservation weight of that residue. Global information is used as in the first-level input.

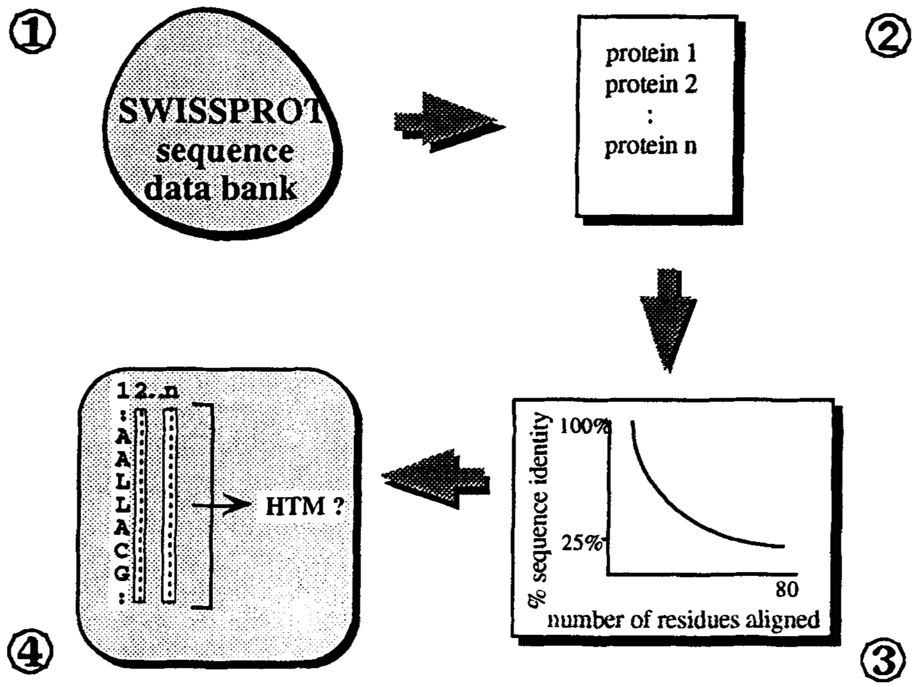


Fig. 3. Generating multiple alignments for the network input. First, for each protein the SWISS-PROT data base of protein sequences (Bairoch & Boeckmann, 1994) was searched for putative homologues with a fast alignment method (FASTA; Pearson & Lipman, 1988; Pearson & Miller, 1992). Second, the list of putative homologues was reexamined with a more sensitive profile-based multiple alignment method (MaxHom; Sander & Schneider, 1991). Third, a length-dependent cutoff for the sequence identity between the search sequence and the aligned ones was applied to distinguish correct hits for homologues from false positives (for more than 80 residues aligned, the cutoff was chosen 25% + 5%; where the "+5%" reflects a safety margin above the line observed to separate correct and false homologues [Sander & Schneider, 1991]). Fourth, a window of 13 adjacent residues was shifted along the protein sequence. Each such window constituted one training or testing example for the neural network.

Table 1. *Prediction accuracy cross-validated on helical transmembrane proteins*[a]

			Overall		Helical transmembrane segments only							
					Per-residue score			Segment-based scores				
Set[b]	Method[c]	N	Q_2	Info	%Obs Q_{TM}	%Prd Q_{TM}	Corr	$\langle L \rangle$	%Obs Sov	%Prd Sov	Nseg[d] over	Nseg under
Set 1	No profiles	69	90	0.45	84	70	0.71	23	90	81	15 6.3%	47 17%
	PHDhtm	**69**	**95**	**0.64**	**91**	**84**	**0.84**	**23**	**96**	**96**	**5** 1.9%	**10** 3.8%
Set 2	PHDhtm	37	95		91		0.85	23				
	Edelman (1993)	37	88		90		0.70	26				
Set 3	Jones et al. (1994)	67									15 4.5%	6 1.9%
Set 4	PHDhtm	28									3-2[e] 1.6%	3 2.3%
	Persson and Argos (1994) Not cross-validated[f]	28									2-3[e] 1.6%	3 2.3%

[a] N, number of proteins used for prediction; Q_2, percentage of correctly predicted residues; Info, information or entropy of prediction (Rost & Sander, 1993b); Q_{TM}, accuracy of predicting transmembrane helices (HTM); %Obs Q_{TM}, correctly predicted residues in HTM as percentage of residues observed in HTM; %Prd Q_{TM}, correctly predicted residues in HTM as percentage of residues predicted as HTM; Corr, Matthews correlation (Matthews, 1975) for residues in HTM; $\langle L \rangle$, average length of predicted HTM (the observed average is $\langle L \rangle = 22$); %Obs Sov, segment overlap for HTM computed as percentage of observed segments (Rost et al., 1994); %Prd Sov, segment overlap for HTM computed as percentage of predicted segments (Rost et al., 1994); Nseg over, number of segments predicted but not observed as HTM; Nseg under, number of segments observed but not predicted as HTM. Bold indicates the reference levels.

[b] Set 1, set of 69 proteins with experimentally well-determined transmembrane helices (see Materials and methods); set 2, set of 37 transmembrane proteins used by Edelman (1993); set 3, set 1 without glra_rat and 2mlt; set 4, set of 28 transmembrane proteins used by Persson and Argos (1994).

[c] No profiles, two-level network system using single sequences as input (R. Casadio et al., submitted); PHDhtm, three-level network system + filter using all information from multiple alignments as input (Fig. 2).

[d] Whenever predicted and observed segments overlapped by at least three residues, the segment was counted as correct (Rost et al., 1993, 1994). A similar measure seems to have been used by others. A more reasonable score is the segment overlap Sov (Rost et al., 1994).

[e] Discrepancy in assigning transmembrane helices for atpi_pea; both methods compared predict five transmembrane helices. In SWISS-PROT only four are annotated; thus, we initially counted our prediction as wrong, whereas Persson and Argos (1994) based their evaluation on the hypothesis that the protein contains five and not four transmembrane helices.

[f] All results except for those in the last row were based on cross-validation tests. Persson and Argos (1994) reported that for their method the results with or without cross-validation analysis are similar and only gave the non-cross-validated results on proteins in their training set.

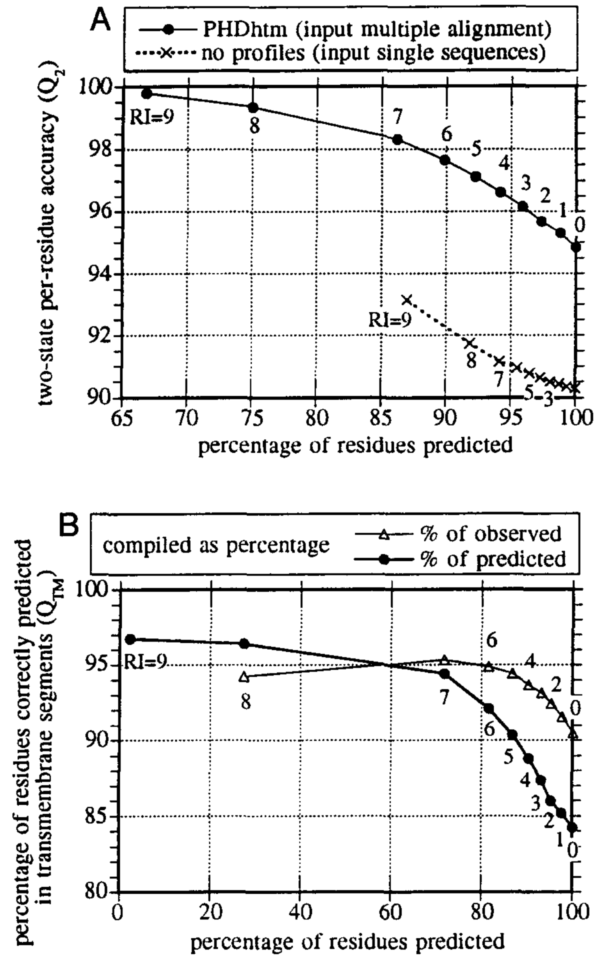


Fig. 4. Reliability of prediction. Reliability index (*RI*) for the prediction was defined as proportional to the difference between the two output units:

$$RI = INTEGER\ (10 \times [\mathrm{out}_{\mathrm{HTM}} - \mathrm{out}_{\mathrm{not\ HTM}}]).$$

The factor 10 scales the reliability index to values 0–9. **A:** Overall two-state per-residue accuracy versus the cumulative percentage of residues with a reliability index $RI \geq n$, $n = 0, \ldots, 9$. Note that $RI \geq 0$ is the rightmost point representing 100% of the predicted residues. Results were averaged over the residues in all 69 transmembrane proteins used for the cross-validation test. A network system that used multiple alignments as input was compared to a network using single sequence information only. For example, 90% of all residues were predicted with $RI \geq 6$. For these, the prediction accuracy for the network using multiple alignment information reached a value of $Q_2 > 97\%$. **B:** Percentage of residues correctly predicted in transmembrane helices versus cumulative percentage of residues predicted in transmembrane helices with a reliability index $RI \geq n$. Results are given as percentages of the number of residues observed in transmembrane helices (open triangles) and as percentages of the number of residues predicted in transmembrane helices (filled circles). For example, about 70% of all residues predicted in transmembrane segments had a reliability index $RI \geq 7$. Ninety-five percent of these were predicted correctly.

jury decision, i.e., the arithmetic average over the output values of balanced and unbalanced networks.

Second-level elongates helices. The effect of the second-level (structure-to-structure) network was to elongate or delete short helical segments. The effect was an increase in the average length of a predicted helical segment from 15 residues for the first level, to 27 residues for the second level (Table 4). In other words, the first-level networks (Fig. 2) yielded an average length for transmembrane segments 5–7 residues shorter than observed; the second-level networks (Fig. 2) resulted in segments up to 13 residues longer than observed. Thus, the second-level networks tended to elongate helices (Table 4).

Final filtering procedure. Short loop regions were often missed by the second network, which tended to elongate helices too much (note that the input window is too narrow to learn a maximal length for transmembrane segments). This drawback was compensated by a relatively straightforward filtering procedure (Materials and methods). Filtering improved the prediction accuracy both in terms of per-residue and segment-based measures for prediction accuracy (Table 4).

Conclusion

Selection of data set. The 3D structure is experimentally known for only five (1prc_H, 1prc_L, 1prc_M, 1brd, 2mlt) of the 69 protein chains used for the cross-validation analysis. This implies that the results ought to be taken with caution. To increase confidence in the results, we deliberately chose proteins for which there is "reliable" experimental evidence about the locations of the transmembrane regions (list taken from Jones et al., 1994), rather than working with a larger data set including less well-known segments.

Improved prediction of transmembrane helices. Using various aspects of evolutionary information improved the overall per-residue accuracy of predicting residues in transmembrane helices by some five percentage points. This improvement could be significant enough to warrant use of the predictions as a starting point for a complete ab initio prediction of 3D structure for transmembrane regions (Baldwin, 1993; Taylor et al., 1994). Our best network system (called PHDhtm) correctly predicted some 94% of all segments and the correct location of some 90% of all residues observed in transmembrane helices. For only 4 of 15 incorrectly predicted (either under-, or overpredicted) segments, the defined reliability index would have led the user to suspect a wrong prediction (Fig. 1).

Prediction for globular proteins sufficiently accurate. The two-level network system using only profiles as input mispredicted less than 2% of globular proteins as containing transmembrane helices (Table 3). An unsatisfactory disadvantage of the most accurate network system PHDhtm was that this error rate was clearly higher (<5%). However, for most practical purposes this rate of false positives is sufficiently low. All transmembrane proteins were predicted to contain at least one transmembrane helix, except for melittin, which would not have been recognized as transmembrane helix even from the crystal structure: the strongly bent helix is split into two short helices by the program assigning the secondary structure automatically from 3D structures (DSSP; Kabsch & Sander, 1983).

Weak point. A rather inconvenient aspect of the method described here is the necessity to apply a filter procedure (Fig. 5) at the end of the prediction. This disadvantage is one of the details that still has to be improved in a more general tool.

```
too short helices

if { $L < 17 \cap RI>7$ (at either end of helix) }-->   elongate helix by one residue
                                                        until $L \geq 17$
if { only one helix predicted }
        if { $L < 17$ }                          -->    cut helix
if { at least 2 helices predicted }
        if { $L < 11$ }                          -->    cut helix

too long helices

if { $L > 35$ }                                  -->    split helix at position L/2
                                                        into two helices of length L/2
if { $L > n \times 22$, n=3,4,... }              -->    split helix into n of length L/n
```

Fig. 5. Filtering the prediction. Output of the third level (jury prediction) was filtered to delete too-short and to split too-long predicted transmembrane helices. Splitting of too-long segments was usually done exactly in the middle of the segment by flipping the prediction for one residue from HTM to not-HTM. Two exceptions were: (1) if there was a residue in a three-residue neighborhood of the central residue with a lower reliability index than that of the central one, then splitting was performed at that residue; (2) if the two residues on both sides of the central residue were predicted with an $RI < 3$, then up to five residues in total were flipped from the state HTM to not-HTM.

Possible improvements of the prediction. There are methods that predict whether or not a loop region is located inside or outside the cell (von Heijne & Gavel, 1988; Nakashima & Nishikawa, 1992; von Heijne, 1992; Sipos & von Heijne, 1993; Jones et al., 1994). Such tools could be used to either complement the network prediction, or directly to train a network to predict transmembrane topology (direction of transmembrane helices with respect to cell).

β-Strand membrane proteins. How can transmembrane segments for β-barrel proteins such as porin be predicted from sequence? Interestingly, the network system trained on water-soluble globular proteins (PHDsec), predicts the β-strands of the membrane protein porin more accurately than the helices of the photoreaction center, bacteriorhodopsin, or the light harvesting complex. The reason may be that the pore of porin is exposed to solvent and thus resembles globular proteins in some respects. The prediction of β-strands, combined with hydrophobicity scales (Eisenberg et al., 1984b) and/or predictions of solvent accessibility (Rost & Sander, 1994b), has been used to infer which of the porin strands may be in contact with lipids. Unfortunately, however, the structures of very few β-strand membrane proteins are known. Thus, training of neural networks, as well as the application of statistical methods, is premature.

3D structure prediction. How can one come closer to the goal of 3D prediction for helical membrane proteins? One way to go from accurate predictions of HTM locations to 3D structure has been indicated by Taylor et al. (1994). Whether or not the network predictions described here, in combination with a prediction of segment orientation relative to the membrane surface, will be useful remains to be shown.

Keeping up with the flow of genome data. All results reported here refer to completely automatic usage of PHDhtm. In some cases, prediction accuracy can certainly be improved by expert knowledge, e.g., by fine tuning the alignment. However, fully automatic use permits the analysis of many proteins, e.g., all open reading frames of complete chromosomes. For example, less than an hour of CPU time (on a SUN SPARC10 workstation) was required for the transmembrane helix prediction of all proteins of yeast chromosome VIII (Johnston et al., 1994), given the multiple sequence alignments. For 59 of the 269 proteins at least two transmembrane helices were predicted (Table 5); for another 27 of the proteins one transmembrane helix was predicted. Given an error rate of 5%, this implies that 20–25% of all yeast VIII proteins were predicted to contain transmembrane helices.

Availability of the network prediction. Predictions of transmembrane helices (as well as secondary structure and solvent accessibility for globular proteins) using the method presented here are provided via an automatic electronic mail server. If you send the sequence of your protein, the server will return a multiple sequence alignment and a prediction of the location of transmembrane helices. For further information, send the word *help* to the Internet address *PredictProtein@EMBL-Heidelberg.DE* by electronic mail, or use the World Wide Web (WWW) site *http://www.embl-heidelberg.de/predictprotein/predictprotein.html.*

Materials and methods

Database

Selection of proteins. We based our analyses on a set of 69 proteins for which experimental information about the location of transmembrane helices is annotated in the SWISS-PROT database (Manoil & Beckwith, 1986; von Heijne & Gavel, 1988; von Heijne, 1992; Sipos & von Heijne, 1993; Jones et al., 1994). This set in particular was chosen to meet three criteria: (1) reliability: the experimental information should be as reliable as possible (Manoil & Beckwith, 1986; von Heijne, 1992); (2) comparability: to enable a comparison to similar methods, the data set should be similar to those used by others; (3) availability: the list (Table 2) was the subset of those proteins used by Jones et al. (1994) that were available in SWISS-PROT when we had started the project (melittin [2mlt] and the glutamic acid receptor [glra_rat, O'Hara et al., 1993] were added). For the few known 3D structures, the location of the transmembrane regions was taken from DSSP (Kabsch & Sander, 1983). The exact locations of the transmembrane helices are often controversial. To enable a straightforward comparison to future methods and for making our results easily reproducible for others, we decided to always use the definitions found in SWISS-PROT (Bairoch & Boeckmann, 1994).

Table 2. *Observed and predicted transmembrane helices for 69 proteins*[a]

Protein	Observed HTM	Predicted HTM
1brd	23-42	24-43
(bacr_halha)	57-76	55-87
	95-114	92-116
	121-140	121-143
	148-167	145-169
	191-210	185-211
	217-236	213-239
1prc_H	12-35	12-31
1prc_M	52-76	43-59
	–	63-78
	111-137	110-130
	143-166	143-170
	198-223	198-223
	260-284	262-292
1prc_L	33-53	21-38
	–	42-58
	84-111	81-103
	116-139	115-146
	171-198	173-196
	226-249	223-255
2mlt	2-10	–
	12-25	–
4f2_human	82-104	82-104
5ht3_mouse	246-272	238-270
	278-296	282-301
	306-324	307-331
	465-484	457-484
a1aa_human	54-79	56-79
	92-117	92-116
	128-150	128-150
	172-196	173-189
	210-233	213-235
	307-331	309-329
	339-363	–
a2aa_human	34-59	32-60
	71-96	69-100
	107-129	106-133
	150-173	151-169
	193-217	196-221
	375-399	375-399
	407-430	405-429
a4_human	700-723	702-722
aa1r_canfa	11-33	12-35
	47-69	39-53
	–	61-74
	81-102	80-110
	124-146	125-144
	177-201	176-206
	236-259	235-261
	268-292	266-291
aa2a_canfa	8-30	10-32
	44-66	40-71
	78-100	77-105
	121-143	122-141
	174-198	174-203
	235-258	234-260
	267-290	266-290
adt_ricpr	34-54	31-46
	68-88	60-87
	93-113	92-115
	148-168	134-148
	–	156-170
	185-205	185-206
adt_ricpr	219-239	217-239
(*continued*)	280-300	271-298
	321-341	322-342
	349-369	348-371
	380-400	377-400
	439-459	444-461
	466-486	469-485
bach_halhm	23-42	24-43
	57-76	55-87
	95-114	92-116
	121-140	121-143
	148-167	145-169
	191-210	185-211
	217-236	213-239
cb21_pea	62-81	69-75
	114-134	115-134
	182-198	184-196
cek2_chick	365-389	371-389
cyoa_ecoli	–	12-24
	51-69	44-66
	93-111	90-109
cyob_ecoli	17-35	–
	58-76	61-77
	102-121	101-131
	144-162	146-158
	195-213	191-212
	232-250	227-252
	277-296	286-302
	320-339	315-335
	348-366	349-368
	382-401	380-401
	410-429	415-440
	457-476	457-470
	494-513	498-519
	588-607	592-608
	614-634	612-626
cyoc_ecoli	32-50	29-50
	67-85	67-85
	102-120	101-116
	143-161	138-162
	185-203	178-202
cyod_ecoli	18-36	20-39
	46-64	45-64
	81-99	80-101
cyoe_ecoli	10-28	12-24
	38-56	44-66
	79-97	90-109
	108-126	109-127
	–	142-158
	–	166-181
	198-216	198-222
	229-247	228-252
	269-287	265-287
edg1_human	47-71	45-72
	79-107	80-107
	122-140	116-145
	160-185	160-180
	202-222	201-227
	256-277	254-282
	294-314	288-312
egfr_human	646-668	648-666
fce2_human	22-47	27-47
glp_pig	63-85	63-84
glpa_human	92-114	91-114
glpc_human	58-81	57-81
glra_rat	539-558	536-557
	585-603	–
	614-632	615-636
	806-826	807-826
gmcr_human	321-346	326-351
gp1b_human	148-172	147-171
gpt_crilo	7-32	12-38
	58-79	59-83
	95-114	96-115
	126-145	127-150
	165-184	157-181
	195-211	187-210
	222-240	224-242
	253-269	249-269
	275-294	277-292
	379-397	379-402
hema_cdvo	35-55	37-58
hema_measi	35-55	37-58
hema_pi4ha	35-59	37-59
hg2a_human	46-72	50-67
iggb_strsp	–	18-32
	–	91-103
	423-443	425-439
il2a_human	241-259	235-258
il2b_human	241-265	236-267
ita5_mouse	356-381	355-383
lacy_ecoli	11-33	11-36
	47-67	46-67
	75-99	75-98
	103-125	104-126
	145-163	148-161
	168-187	169-187
	212-234	219-238
	260-281	265-288
	291-310	294-314
	315-334	320-337
	347-366	343-371
	380-399	377-400
lech_human	40-60	40-59
leci_mouse	40-60	40-59
lep_ecoli	4-22	4-23
	58-76	63-82
magl_mouse	517-536	515-534
malf_ecoli	17-35	21-35
	40-58	43-58
	73-91	71-93
	277-295	278-306
	319-337	318-339
	371-389	370-390
	418-436	418-444
	486-504	486-505
motb_ecoli	28-49	30-51
mprd_human	186-210	185-211
myp0_human	–	14-31
	154-179	155-183
ngfr_human	251-272	253-272

(*continued*)

Table 2. *Continued*

Protein	Observed HTM	Predicted HTM
nep_human	28-50	30-49
oppb_salty	10-30	10-29
	100-121	96-120
	138-158	130-162
	173-190	168-193
	227-250	228-259
	272-293	273-298
oppc_salty	38-59	39-59
	102-122	98-126
	140-160	141-158
	164-180	166-182
	216-236	210-225
	–	232-248
	268-290	268-289
ops1_calvi	48-72	47-75
	85-110	85-110
	125-144	116-145
	164-187	162-187
	212-239	212-239
	275-298	275-298
	306-330	306-329
ops2_drome	57-81	55-84
	94-119	94-118
	134-153	124-153
	173-196	171-196
	221-248	221-248
	284-307	284-307
	315-339	315-338
ops3_drome	58-82	57-85
	95-119	95-119
ops3_drome (*continued*)	134-152	125-153
	172-196	169-194
	221-248	221-248
	285-308	285-308
	317-341	317-340
ops4_drome	54-78	53-81
	91-113	91-115
	130-149	121-150
	168-192	166-191
	217-244	217-244
	281-304	281-304
	313-337	313-336
opsb_human	34-58	33-59
	71-96	71-100
	111-130	112-135
	150-173	149-173
	200-227	200-227
	250-272	251-275
	282-306	281-306
opsd_bovin	37-61	36-62
	74-99	74-104
	114-133	115-139
	153-176	152-176
	203-230	203-230
	252-276	253-279
	285-309	285-309
opsg_human	53-77	52-78
	90-115	90-120
	130-149	131-155
	169-192	168-192
opsg_human (*continued*)	219-246	219-245
	269-292	269-295
	301-325	301-325
opsr_human	53-77	52-78
	90-115	90-119
	130-149	131-155
	169-192	168-192
	219-246	219-245
	269-292	270-295
	301-325	301-325
pigr_human	621-643	624-643
pt2m_ecoli	25-44	20-42
	51-69	54-65
	135-154	133-156
	166-184	167-181
	–	249-262
	274-291	270-283
	314-333	312-332
sece_ecoli	19-36	20-34
	45-63	42-62
	93-111	93-123
suis_human	13-32	12-33
tcb1_rabit	292-313	285-312
trbm_human	516-539	515-536
trsr_human	63-88	67-86
vmt2_iaann	25-42	27-51
vnb_inbbe	19-40	19-42

[a] For the 69 transmembrane proteins used for cross-validation, the following data are listed: (1) the protein name, given by the SWISS-PROT identifier (Bairoch & Boeckmann, 1994); if the 3D structure is known, then the PDB code plus chain identifier is used (Bernstein et al., 1977; Kabsch & Sander, 1983); (2) the positions for the transmembrane helices observed (=SWISS-PROT documentation, or DSSP [Kabsch & Sander, 1983]), counted from the first residue in SWISS-PROT or DSSP; and (3) the cross-validated prediction by the network system PHDhtm. Except for 2mlt and glra_rat, the list comprises a subset of the proteins used by David Jones (Jones et al., 1994) and Gunnar von Heijne (von Heijne & Gavel, 1988; von Heijne, 1992; Sipos & von Heijne, 1993).

Generation of multiple alignments. For each of the initial 69 proteins, a multiple sequence alignment was generated using the program MaxHom (Sander & Schneider, 1991; Fig. 3). All sequences from SWISS-PROT with a sequence identity above a length-dependent cut-off were included in the alignment (Sander & Schneider, 1991), assuming that this is valid not only for globular but also for membrane proteins.

Cross-validation test. The set of 69 transmembrane proteins (Table 2) was divided into 52 proteins used for training and 17 used for testing the method. This was repeated five times (five-fold cross-validation), until each protein had been in a test set once. The sets were chosen such that no protein in the multiple alignments used for testing had more than 25% sequence identity to any protein in the multiple alignments of the training set. All results reported are averages over proteins in various test sets.

Neural network system

First level: Sequence-to-structure. The principles of neural networks for secondary structure prediction (Fariselli et al., 1993; Rost & Sander, 1993a) and of coding multiple sequence information (Rost & Sander, 1993b, 1994a, 1994b) are described in detail elsewhere. Here, only some basic concepts will be recapitulated and details regarding the application to transmembrane helices will be introduced.

Input to the first-level network consisted of two contributions, (1) one local in sequence, i.e., taken from a window of 13 adjacent residues; and (2) another global in sequence, i.e., compiled from the whole protein (Fig. 2). (1) The local information computed for each residue in the window was the frequency of occurrence of each amino acid at that position in the multiple alignment, the number of insertions and deletions in the alignment for that residue, and a position-specific conservation weight (Fig. 2). (2) As global information, we used the amino acid composition and length of the protein and, furthermore, the distance (number of residues) of the first residue in the window of 13 adjacent residues from the protein begin (N-term), and the distance of the last residue in the window to the protein end (C-term).

Output of the first-level network was two units, one representing examples with the central residue of the window in a

Table 3. *Prediction accuracy on globular proteins (negative control)*[a]

Method	Number of globular proteins used	Number of proteins predicted with HTM	Number of HTM segments longer than 16 residues	% False classifications
No profiles	278	18	7	6.5%
Profiles only	278	5	4	1.8%
PHDhtm	278	12	7	4.3%
Jones et al. (1994)	155	5	–	3.2%
Edelman (1993)	14	3	–	21.4%

[a] Abbreviations for methods as in Table 1 and Table 4. We considered a globular protein to be mispredicted if either at least two transmembrane segments are predicted with more than 10 residues, or at least one with more than 17 residues. Results from Edelman (1993) and Jones et al. (1994) were taken from the literature.

transmembrane helix; the other representing examples with the central residue not in transmembrane helices (Fig. 2).

Balanced and unbalanced training. Training was performed with the usual gradient descent (also known as back-propagation [Rumelhart et al., 1986]):

$$\Delta J_{ij}(t+1) = \epsilon \frac{E(t)}{J_{ij}(t)} + \alpha \Delta J_{ij}(t-1),$$

where t is the algorithmic time step (i.e., change of all connections for one pattern), E is the error, given by the difference between actual network output and the desired output (i.e., the value observed for the central residue); J_{ij} is the connection from unit j to unit i on the next layer (input to hidden, hidden to output); ϵ is the learning speed, chosen here to be 0.01; and α the momentum term (permitting uphill moves) chosen here to be 0.2. Two modes were used. First, unbalanced training: at each time step of the error minimization one pattern was chosen at random from the training set, and all connections of the network were changed. Second, balanced training: at each time step of the error minimization (Equation 1), one pattern from the class "transmembrane helix" and one from the class "not transmembrane helix" was used to change all connections.

Table 4. *Analysis of the performance for each element of the network system*[a]

			Overall		Transmembrane helices only					
					Per-residue score			Segment-based scores		
Set	Method[b]	System levels[c]	Q_2	Info	%Obs Q_{TM}	%Prd Q_{TM}	Corr	⟨L⟩	%Obs Sov	%Prd Sov
Set 5	No profiles	2 + filter	90	0.45	84	70	0.71	23	90	81
	Profiles only	2 + filter	94	0.56	86	82	0.80	23	93	90
	PHDhtm	3 + filter	**95**	**0.65**	**91**	**84**	**0.85**	**23**	**96**	**96**
Set 1	First unbalanced	1	93	0.52	78	81	0.75	15	84	80
	First balanced	1	91	0.53	91	71	0.76	17	80	72
	First unbalanced–second unbalanced	2	93	0.52	83	80	0.77	22	88	83
	First balanced–second unbalanced	2	93	0.52	83	80	0.77	22	88	83
	First unbalanced–second balanced	2	91	0.55	91	69	0.75	36	71	63
	First balanced–second balanced	2	93	0.58	93	75	0.79	29	80	75
	Jury over four networks	3	91	0.58	94	69	0.75	36	71	63
	PHDhtm	3 + filter	**95**	**0.64**	**91**	**84**	**0.84**	**23**	**96**	**96**

[a] See Table 1 for abbreviations of measures. Bold indicates the reference levels for each set.

[b] PHDhtm, three-level network system + filter using all information from multiple alignments as input (Fig. 2); No profiles, two-level network system using single sequences as input (R. Casadio et al., submitted); Profiles only, same as before, but using evolutionary profiles (and no further information derived from the multiple alignment) as input; First unbalanced, first-level network with unbalanced training (see Materials and methods); First balanced, first-level network with balanced training (see Materials and methods); First x-second y, a second-level network with y (balanced or unbalanced) training that uses as input the prediction from a first-level network with x (balanced or unbalanced) training; Jury over four networks, arithmetic average over the four different second-level networks given above.

[c] Levels of the network system used (Fig. 2): 1, only first level; 2, first and second level; 3, jury average over different second-level networks (see Materials and methods); filter, application of the filtering procedure (Fig. 5). Set 1 contains 69 transmembrane proteins (see Materials and methods). Set 5 is the subset of set 1 without the PDB proteins 2mlt, 1prc (chains H, L, M), and 1brd.

Table 5. *Prediction of transmembrane helices for yeast chromosome VIII*[a]

Identifier	Nres[b]	Nali[b]	Locations of predicted segments				Nhtm[b]
YHL040c	627	5	75-88	116-127	141-157	173-190	
			205-216	231-252	285-308	326-342	
			363-387	404-418	429-441	458-477	
			568-581				13
YHL047c	637	5	70-83	111-122	136-152	168-185	
			200-211	226-247	280-303	321-337	
			358-382	400-413	425-436	453-473	
			563-576				13
YHR092c	560	21	70-87	124-139	152-171	179-196	
			215-226	247-261	369-385	400-413	
			435-459	474-492	500-518		11
YHR096c	592	18	85-101	138-154	167-186	194-212	
			230-241	262-276	385-400	415-428	
			450-475	489-507	515-533		11
YHR094c	570	17	64-80	118-133	146-165	173-191	
			209-220	241-255	363-379	394-407	
			429-453	468-486	494-512		
YHR026w	213	18	20-37	56-80	94-122	145-168	
			180-205				5
YHR002w	357	8	37-53	102-115	141-153	201-227	
			271-281				5
YHL048w	381	4	39-62	70-93	233-252	260-277	4
YHR190w	444	4	272-283	295-310	425-440		3
YHR129c	384	258	137-153	349-360			2
YHR005c	472	153	337-347	377-387			2
YHR183w	489	39	360-371	418-429			2
YHR046c	295	7	103-117	201-216			2
YHR176w	373	6	262-272	338-351			2
YHR039c	644	5	49-66	247-264			2
YHL011c	320	22	73-92				1
YHR028c	818	8	26-44				1
YHR007c	530	7	25-47				1
YHR037w	575	4	209-227				1
YHL016c	735	1	17-33	91-108	137-153	167-186	
			193-213	256-266	287-311	339-350	
			358-375	402-421	429-450	458-476	
			500-516	620-642	651-674		15
YHL035c	1,592	1	33-48	172-187	201-217	229-239	
			335-357	378-395	465-486	490-510	
			574-591	977-998	1,042-1,058	1,120-1,137	
			1,141-1,158	1,226-1,247	1,255-1,274		15
YHL036w	546	1	69-92	100-122	149-171	187-203	
			211-235	261-273	298-315	345-367	
			398-413	433-445	461-477	492-519	12
YHR048w	514	1	75-91	112-126	143-160	168-184	
			197-221	229-249	308-334	343-364	
			390-407	415-438	478-498		11
YHR050w	549	1	92-106	135-156	164-181	199-218	
			246-257	309-333	361-376	409-423	
			434-451	518-538			10
YHR123w	391	2	40-67	123-156	177-199	218-235	
			267-286	294-312	320-342	350-372	8
YHL003c	411	3	82-100	133-160	181-198	216-238	
			256-288	303-319	353-383		7
YHL017w	532	2	194-212	227-243	260-290	307-318	
			331-353	376-399	420-438		7
YHR050w	549	1	92-106	135-156	164-181	199-218	
			246-257	309-333	361-376	409-423	

[a] As a typical example for the application of the method and as an independent test of the predictive power of the method, we predicted the transmembrane helices for all proteins from the complete yeast chromosome VIII (Johnston et al., 1994). For 59 proteins (of 269), two or more transmembrane helices were predicted. Proteins are labeled by the identifier used in Johnston et al. (1994). Shown are the predictions only for those proteins for which sufficient alignment information was available (P. Bork, C. Ouzounis, & C. Sander, manuscript in prep.) or which were predicted to have more than six transmembrane segments. In some cases, confirmation of the correctness of the prediction comes from detailed sequence analysis (Johnston et al., 1994; P. Bork, C. Ouzounis, & C. Sander, unpubl.): the likely function identified on the basis of sequence similarity to proteins of known function is consistent with the presence of HTM regions. Examples are: YHR026w, an ATPase; YHR048w, a resistance protein, probably works by pumping substances out of the cell through a membrane pore; YHR050w/92c/94c/96c, potential transporters; YHR190w, farnesyltransferase; YHR123w, phosphor transferase; YHR005c, G-protein α subunit; YHR183w/39c, dehydrogenase.

[b] Nres, length of protein; Nali, number of sequences in the multiple alignment ("1" means that the prediction is based on a single sequence only); Nhtm, predicted number of transmembrane segments.

Network parameters. All units were connected to all those on the next layer (input to hidden, hidden to output). Network parameters such as criterion to terminate the training procedure, number of hidden units, training speed (ϵ in Equation 1), and momentum term (α in Equation 1) were chosen arbitrarily based on our experience with secondary structure prediction for globular proteins. In other words, these parameters were not influenced by the test set. Training was stopped when the training set had been learned to an accuracy of 93% for the first- and of 95% for the second-level network. As for the number of hidden units, we started arbitrarily with 3 hidden units for the first level of network and increased the number for the second-level network to 15 because training too often ended in local minima.

Second level: Structure to structure. The input to the second-level network consisted—as for the first-level—of a contribution local in sequence and a contribution global in sequence (Fig. 2). (1) For each residue in the input window, the local input were the values of the two output units of the first-level network and the conservation weight. (2) The global input information was the same as for the first-level network. The output of the second-level network—as for the first—consisted of two units for the central residue either being in a transmembrane helix or not.

Third level: Jury decision. To find a compromise between networks with balanced and those with unbalanced training, a final jury decision was performed (effectively a compromise between over- and underprediction, Results). The jury decision was a simple arithmetic average over four differently trained networks: all combinations (2 × 2) of first-level network with balanced and unbalanced training, and with balanced or unbalanced training of second-level network. Final prediction was assigned to the unit with maximal output value ("winner takes all").

Fourth level: Filtering the prediction. In contrast to earlier prediction methods (Jones et al., 1992; von Heijne, 1992; Persson & Argos, 1994), which explicitly fix the length of predicted transmembrane segments to typically 17-25 residues, the second-level network occasionally resulted in transmembrane helices that were either too short or too long. This was corrected by a nonoptimized filter that was guided by the experiences of previous work (von Heijne, 1986, 1992; von Heijne & Gavel, 1988; Sipos & von Heijne, 1993; Jones et al., 1994; R. Casadio et al., submitted).

Too long helices were either split in the middle into two shorter helices or were shortened (Fig. 5). Too short helices were either elongated or deleted. All these decisions (split or shorten; elongate or delete) were based both on the strength of the prediction (reliability index, Fig. 2) and on the length of the predicted transmembrane helix (Fig. 5).

Acknowledgments

We are grateful to Reinhard Schneider (EMBL, Heidelberg) for providing the latest version of the alignment program MaxHom; Chiara Taroni (Bologna) and Mario Compiani (Camerino) for helpful discussions; David Jones (London) for help with the data set; Gunnar von Heijne (Huddinge) for motivating discussions; and Christos Ouzounis (EMBL) for providing the multiple alignments for yeast VIII. We thank the two referees, who helped improve the text by their detailed criticism. Last, but not least, we thank all those who deposit experimental results in public databases.

References

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

Bairoch A, Boeckmann B. 1994. The SWISS-PROT protein sequence data bank: Current status. *Nucleic Acids Res 22*:3578-3580.

Baldwin JM. 1993. The probable arrangement of the helices in G protein-coupled receptors. *EMBO J 12*:1693-1703.

Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, 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.

Cornette JL, Cease KB, Margalit H, Spouge JL, Berzofsky JA, DeLisi C. 1987. Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *J Mol Biol 195*:659-685.

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

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, Mii 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.

Edelman J. 1993. Quadratic minimization of predictors for protein secondary structure: Application to transmembrane α-helices. *J Mol Biol 232*: 165-191.

Eisenberg D, Schwartz E, Komaromy M, Wall R. 1984a. Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J Mol Biol 179*:125-142.

Eisenberg D, Weiss RM, Terwilliger TC. 1984b. The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc Natl Acad Sci USA 81*:140-144.

Engelman DM, Steitz TA, Goldman A. 1986. Identifying nonpolar transbilayer helices 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.

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.

Johnston M, et al. [35 authors]. 1994. Complete nucleotide sequence of *Saccharomyces cerevisiae* chromosome VIII. *Science 265*:2077-2082.

Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. *CABIOS 8*:275-282.

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 recognition of hydrogen bonded and geometrical features. *Biopolymers 22*:2577-2637.

Kühlbrandt W, Wang DN, Fujiyoshi Y. 1994. Atomic model of plant light-harvesting complex by electron crystallography. *Nature 367*:614-621.

Kyte J, Doolittle RF. 1982. A simple method for displaying the hydropathic character of a protein. *J Mol Biol 157*:105-132.

Landolt-Marticorena C, Williams KA, Deber CM, Reithmeier RAF. 1992. Non-random distribution of amino acids in the transmembrane segments of human type I single span membrane proteins. *J Mol Biol 229*:602-608.

Lattman EE. 1994. Protein crystallography for all. *Proteins Struct Funct Genet 18*:103-106.

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.

Nakashima H, Nishikawa K. 1992. The amino acid composition is different between the cytoplasmic and extracellular sides in membrane proteins. *FEBS Lett 303*:141-146.

O'Hara PJ, Sheppard PO, Thφgersen H, Venezia D, Haldeman BA, McGrane V, Houamed KM, Thomsen C, Gilbert TL, Mulvihill ER. 1993. The ligand-binding domain in metabotropic glutamate receptors is related to bacterial periplasmic binding proteins. *Neuron 11*:41-52.

Oliver S, et al. [152 authors]. 1992. The complete DNA sequence of yeast chromosome III. *Nature 357*:38-46.

Pearson WR, Lipman DJ. 1988. Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA 85*:2444-2448.

Pearson WR, Miller W. 1992. Dynamic programming algorithms for biological sequence comparison. *Methods Enzymol 210*:575-601.

Persson B, Argos P. 1994. Prediction of transmembrane segments in proteins utilising multiple sequence alignments. *J Mol Biol 237*:182-192.

Rost B, Sander C. 1993a. Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proc Natl Acad Sci USA 90*:7558–7562.

Rost B, Sander C. 1993b. Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol 232*:584–599.

Rost B, Sander C. 1993c. Secondary structure prediction of all-helical proteins in two states. *Protein Eng 6*:831–836.

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.

Rost B, Sander C, Schneider R. 1993. Progress in protein structure prediction? *Trends Biochem Sci 18*:120–123.

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 representations by back-propagating error. *Nature 323*:533–536.

Sander C, Schneider R. 1991. Database of homology-derived structures and the structural meaning of sequence alignment. *Proteins Struct Funct Genet 9*:56–68.

Sander C, Schneider R. 1994. The HSSP database of protein structure-sequence alignments. *Nucleic Acids Res 22*:3597–3599.

Sipos L, von Heijne G. 1993. Predicting the topology of eukaryotic membrane proteins. *Eur J Biochem 213*:1333–1340.

Taylor WR, Jones DT, Green NM. 1994. A method for α-helical integral membrane protein fold prediction. *Proteins Struct Funct Genet 18*: 281–294.

von Heijne G. 1981. Membrane proteins—The amino acid composition of membrane-penetrating segments. *Eur J Biochem 120*:275–278.

von Heijne G. 1986. A new method for predicting signal sequence cleavage sites. *Nucleic Acids Res 14*:4683–4690.

von Heijne G. 1991. Computer analysis of DNA and protein sequences. *Eur J Biochem 199*:253–256.

von Heijne G. 1992. Membrane protein structure prediction. *J Mol Biol 225*:487–494.

von Heijne G, Gavel Y. 1988. Topogenic signals in integral membrane proteins. *Eur J Biochem 174*:671–678.

Wang DN, Kühlbrandt W, Sarabiah V, Reithmeier RAF. 1993. Two-dimensional structure of the membrane domain of human Band 3, the anion transport protein of erythrocyte membrane. *EMBO J 12*:2233–2239.

Weiss MS, Schulz GE. 1992. Structure of porin refined at 1.8 Å resolution. *J Mol Biol 227*:493–509.