

Refining neural network predictions for helical transmembrane proteins by dynamic programming

Burkhard Rost [◇],
EMBL, 69012 Heidelberg, Germany
EBI, Hinxton, Cambridge CB10 1RQ, U.K.
Rost@embl-heidelberg.de

Rita Casadio, and Piero Fariselli
Lab. of Biophysics, Dep. of Biology
Univ. of Bologna, 40126 Bologna, Italy
Casadio@kaiser.alma.unibo.it

Abstract

For transmembrane proteins experimental determination of three-dimensional structure is problematic. However, membrane proteins have important impact for molecular biology in general, and for drug design in particular. Thus, prediction methods are needed. Here we introduce a method that started from the output of the profile-based neural network system PHDhtm (Rost, et al. 1995). Instead of choosing the neural network output unit with maximal value as prediction, we implemented a dynamic programming-like refinement procedure that aimed at producing the best model for all transmembrane helices compatible with the neural network output. The refined prediction was used successfully to predict transmembrane topology based on an empirical rule for the charge difference between extra- and intra-cytoplasmic regions (positive-inside rule). Preliminary results suggest that the refinement was clearly superior to the initial neural network system; and that the method predicted all transmembrane helices correctly for more proteins than a previously applied empirical filter. The resulting accuracy in predicting topology was better than 80%. Although a more thorough evaluation of the method on a larger data set will be required, the results compared favourably with alternative methods. The results reflected the strength of the refinement procedure which was the successful incorporation of global information: whereas the residue preferences output by the neural network were derived from stretches of 17 adjacent residues, the refinement procedure involved constraints on the level of the entire protein.

Introduction

Given the rapid advance of large-scale gene-sequencing projects (Fleischmann, et al. 1995), most protein sequences of key organisms will probably be known in a few years'

[◇] Corresponding author.

Abbreviations: **3D**, three-dimensional; **HTM**, transmembrane helix (also abbreviated by H in figures, L used for non-transmembrane regions); **PDB**, Protein Data Bank of experimentally determined 3D structures of proteins; **PHDhtm**, Profile based neural network prediction of helical transmembrane regions; **SWISS-PROT**, data base of known protein sequences. Currently, the gap between the number of known protein sequences and protein three-dimensional (3D) structures is still increasing (Rost 1995). However, experimental structure determination is becoming more of a routine (Lattman 1994). Thus, within the next decades we may know the three-dimensional (3D) structure for most proteins. Even in such an optimistic scenario, experimental information is likely to remain scarce for integral membrane proteins. These challenge experimental structure determination: X-ray crystallography is problematic as the hydrophobic molecules hardly crystallise; and for nuclear magnetic resonance (NMR) spectroscopy transmembrane proteins are usually too long. Today, 3D structure is known for only six membrane proteins (Rost, et al. 1995). Can theory predict structural aspects for integral membrane proteins?

Theoretical predictions for helical transmembrane proteins. The prediction of structural aspects is simpler for membrane proteins than for globular proteins as the lipid bilayer imposes strong constraints on the degrees of freedom of 3D structure (Taylor, et al. 1994). Most prediction tools focus on helical transmembrane proteins for which some experimental information about 3D structure is

available (Manoil & Beckwith 1986, Hennessey & Broome-Smith 1993). Methods have been designed to predict the locations of transmembrane helices

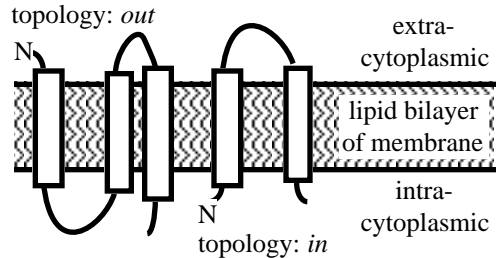


Fig. 1. Definition of topology. In one of the two known classes of membrane proteins, typically apolar helices are embedded in the lipid bilayer oriented perpendicular to the membrane surface. The helices can be regarded as more or less rigid cylinders. 'Topology' is defined as *out* when the N-term (first residue) starts on the extra-cytoplasmic side and as *in* if it starts on the intra-cytoplasmic side. (Kyte & Doolittle 1982, Engelman, et al. 1986, von Heijne 1986, von Heijne 1992, Edelman 1993, Jones, et al. 1994, Persson & Argos 1994, Donnelly & Findlay 1995), and the orientation of transmembrane helices with respect to themembrane (dubbed topology, Fig. 1; (von Heijne 1986, Hartmann, et al. 1989, von Heijne 1989, Sipos & von Heijne 1993, Jones, et al. 1994, Casadio & Fariselli 1996). If the transmembrane helix locations and topology are known with sufficient accuracy, 3D structure can be successfully predicted for the membrane spanning segments by an exhaustive search of the entire structure space (Taylor, et al. 1994).

Further improvement of prediction accuracy necessary?. There are two incentives to aim at improving the accuracy of predicting transmembrane helix locations. (1) In general, current prediction methods are probably not accurate enough to be used for the 3D prediction proposed by Taylor and colleagues (Taylor, et al. 1994). (2) A simple means to predict topology (Fig. 1) is the positive-inside rule: positively charged residues are abundant in periplasmic loops of membrane proteins (von Heijne 1986, von Heijne & Gavel 1988, Hartmann, et al. 1989, Boyd & Beckwith 1990, von Heijne 1992). To successfully apply this rule, a correct prediction of the non-transmembrane regions is required. Would a

better prediction of topology require to predict more accurately each residue (per-residue

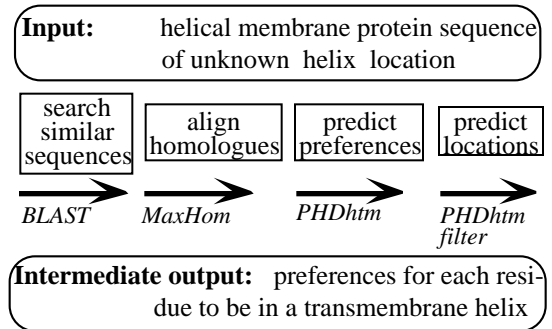


Fig. 2. Neural network predictions for transmembrane helix preferences (PHDhtm). Input was a protein sequence; output were the locations of transmembrane helices. (1) Similar sequences were searched for in SWISS-PROT (Bairoch & Boeckmann 1994) by the fast alignment algorithm BLAST (Altschul, et al. 1990). (2) The resulting sequences were re-aligned by the multiple alignment algorithm MAXHOM (Sander & Schneider 1991). (3) Profiles derived from the alignment were input into the neural network system PHDhtm (Rost, et al. 1995). (4) The final output of PHDhtm were the locations of transmembrane helices assigned according to the output unit with maximal value (winner-takes-all).

accuracy), or would it suffice to predict more trans-membrane helices correctly (per-segment accuracy; (Rost, et al. 1994, Rost, et al. 1995)?

Here we introduce a conceptually simple method that started from the output of the profile-based neural network system PHDhtm (Fig. 2). The preferences for residues to be in transmembrane helices output from the neural network system were post-processed by a dynamic programming-like algorithm. A similar algorithm has been used before to improve statistics-based predictions for transmembrane helices (Jones, et al. 1994). The algorithm constitutes an example for a problem-specific improvement of neural network output. The resulting predictions were used to predict topology based on charge differences between extra- and intra-cytoplasmic residues (Fig. 1). The method is described in detail; and some preliminary results are given.

Materials and methods

Database and evaluation of method

Selection of proteins. We based our analyses on a set of 68 proteins for which experimental information about the locations of transmembrane helices is annotated in the SWISS-PROT database (von Heijne 1992, Bairoch & Boeckmann 1994). The locations of transmembrane helices used are listed in our previous work (Rost, et al. 1995). The observed topology was taken from SWISS-PROT or (Jones, et al. 1994)(Jones, et al. 1994; names listed in Table 2). Note that the set was the same as used previously (Rost, et al. 1995), except for melittin, that was excluded for the purpose of topology prediction.

Cross-validation test. For the prediction of transmembrane propensities by the neural network system (PHDhtm; (Rost, et al. 1995), the set of 68 transmembrane proteins (Table 2) was divided into 55 proteins used for training the networks and 13 used for deriving the propensities (test set). This was repeated five times (five-fold cross-validation) such that each protein was in one test set. The sets were chosen such that no protein in the multiple alignments used for training the networks had more than 25% pairwise sequence identity to any protein in the multiple alignments of the proteins for which the propensities were predicted. The cross-validation procedure yields estimates for prediction accuracy that are likely to hold for proteins of yet unknown topology (Rost & Sander 1993, Rost & Sander 1995, Rost & Sander 1996).

Measuring prediction accuracy. We compiled results for per-residue scores capturing the accuracy of predicting HTM placement (Table 1); and per-segment scores capturing the accuracy of predicting entire HTM's correctly. Due to the length of HTM's and the prediction accuracy, the definition of segment-based scores was straightforward (in contrast to the case of secondary structure predictions for globular proteins (Rost, et al. 1994)). We regarded a transmembrane helix to be predicted correctly, if the overlap between observed and predicted helix was at least five residues (Table 1). Furthermore, we compiled the percentage of proteins for which all HTM's were correctly

predicted (Q_M) and the percentage of proteins for which topology was correctly predicted (Q_T). Note: the topology prediction was regarded as correct if and only if all HTM's and the orientation of the HTM's with respect to the membrane were correctly predicted.

Post-processing neural network output by dynamic programming

Neural network predictions of HTM propensities. Our previously published prediction tool PHDhtm (Rost, et al. 1995) uses multiple sequence alignments as input to a neural network system (Fig. 2). The network output are two values for each residue, describing the propensity of that residue to be in a transmembrane helix (H) or to be in a region outside of the lipid bilayer (L). We used two levels of networks: a first sequence-to-structure level and a

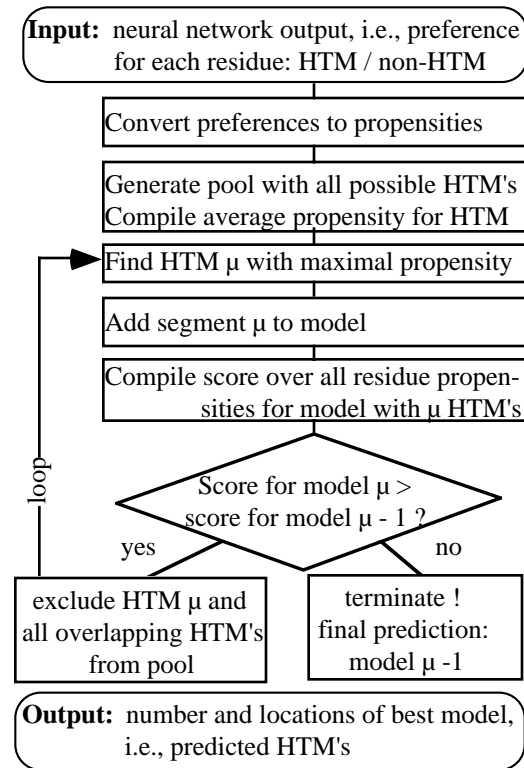


Fig. 3. Flow chart for refinement algorithm. Input to the refinement are the PHDhtm preferences for each residue to be in the HTM or in the non-HTM state. Output are the predictions of the number and locations of all segments in the protein.

second structure-to-structure level (Rost 1996). The second level captures correlations between adjacent residues, i.e., learns to regard transmembrane helices as objects of a minimal length. The effect of the second level is an increase in the average length of a predicted helix. However, the second level frequently predicts too long helices. We previously corrected this by introducing an empirical filter (PHDhtm filter) that relied crucially on choosing free parameters (Rost, et al. 1995).

Finding the optimal path through all predicted propensities (dynamic programming). The simplest way to obtain predictions for HTM locations from propensities is to predict each residue to be in the state (H or L) with largest propensity. An alternative for generating predictions from propensities is an algorithm similar to dynamic programming which has been investigated for predicting transmembrane topology by Jones, Taylor and Thornton (Jones, et al. 1994). The main purpose of this algorithm, as we implemented it, was to generate the optimal placements of HTM segments compatible with the neural network output. The network residue preferences were derived from information largely local in sequence (input to neural network: profiles for 17 adjacent residues). The refinement involved using information about the entire protein, and hence, the procedure - to some extent - made use of global information. The refinement was implemented by the following four steps (Fig. 3).

(1) *Converting neural network output to propensities.* The residue preferences output by PHDhtm were converted to residue propensities p , normalised to sum to one for each residue:

$$p_i^H = \frac{o_i^H}{o_i^H + o_i^L} \text{ and } p_i^L = 1 - p_i^H, \text{ for } i = 1, \dots, N_{\text{res}} \quad (1)$$

where N_{res} was the number of residues (protein length), o_i^H and o_i^L the values of the network output unit coding for the HTM and the non-HTM state of residue i .

(2) *Compiling average propensity for all possible transmembrane helices.* The average propensity for all possible HTM's T were compiled by summing the HTM propensities over all residues of T :

$$T_j = \frac{1}{L_m} \sum_{k=i}^{i+L_m-1} p_k^H, \quad (2)$$

for $i = 1, \dots, N_{\text{res}}$ and $m = L^{\text{min}}, \dots, L^{\text{max}}$ where L_m described the length of the j -th HTM T_j ; L^{min} and L^{max} were the minimal and maximal length allowed for HTM's (eq. 7). The index j labelled all possible HTM's, running from 1 to the number of all possible HTM's N_{seg} :
 $N \setminus \setminus \text{DO3}(\text{seg}) = (N \setminus \setminus \text{DO3}(\text{seg}) -$

$$L^{\text{max}+1}) \times (L^{\text{max}} - L^{\text{min}} + 1) + \sum_{k=1}^{L^{\text{max}} - L^{\text{min}}} k \quad (3)$$

Note: N_{seg} is usually much larger than the number of residues as the index j labels all constructable HTM's, e.g. for HTM's between 5-8 residues and a protein of 8 residues the following 10 HTM's could be constructed (residue number of first and last residue in HTM):

1-5, 1-6, 1-7, 1-8, 2-6, 2-7, 2-8, 3-7, 3-8, 4-8.

(3) *Building optimal path by successively adding strongest HTM.* The dynamic programming aspect of the algorithm was implemented by starting from a model M_0 with no HTM and iteratively adding the HTM with maximal value T (eq. 2):

$$M_\mu = M_{\mu-1} + \max_{j=1, \dots, N_{\text{seg}}} T_j \quad (4)$$

where, e.g., $M_{\mu-1}$ described the model which predicted the protein to contain a single HTM. The compilation of the maximum over all HTM's T was object to two constraints. (i) All segments that overlap with the segment added at iteration step μ were excluded for the following steps (this guaranteed that at any step μ the resulting model was the best model consistent with the assumption that the protein consists of exactly μ HTM's). (ii) The regions excluded were extended by a minimal loop region of length four residues to separate two adjacent HTM's (eq. 7). The score for the model with μ HTM's was defined by the sum over all per-residue properties:

$$P_\mu = \frac{1}{N_{\text{res}}} \sum_{k=1}^{N_{\text{res}}} p_k^H \delta_k^H + p_k^L \delta_k^L \quad (5)$$

with $\delta_k^H = \begin{cases} 1, & \text{if residue } k \text{ in a HTM} \\ 0, & \text{else} \end{cases}$, and $\delta_k^L = 1 - \delta_k^H$

(4) *Selecting model with best score.* Finally, the model with maximal score was chosen as prediction. However, we kept all 'sub-optimal' models to enable users to focus on the strongest helices predicted. Thus, expert information about the protein could be used to favour a model that has a lower.

Defining the reliability of the best model. The refinement procedure results in various possible models (i.e. different predictions for the number of HTM's). In particular, the difference between the score of the best and the second best model rendered a measure for the reliability of the final best model, i.e., for the reliability that the predicted protein had, say μ' , HTM's:

$$Ri_M = P_{\mu'} - \max_{\mu \neq \mu'} P_{\mu} \quad (6)$$

Choosing free parameters. The dynamic programming-like refinement procedure was less arbitrary than was the previously used filter (Rost, et al. 1995), since only three parameters had to be chosen. The choices were made by optimising the performance on a subset of 10 proteins. For the minimal and maximal length of HTM's, and the minimal length of non-transmembrane regions inserted between two HTM's, we used:

$$L^{\min} = 18, \quad L^{\max} = 25, \quad L^{\text{loop}} = 4 \quad (7)$$

Using the refined model to predict transmembrane topology

Verifying the positive-inside rule. Gunnar von Heijne established that integral membrane proteins of various species contain more positively charged residues (Arginine and Lysine) on the cytoplasmic side of transmembrane helices than on the periplasmic side (von Heijne & Gavel 1988, von Heijne 1992). Before applying this rule to our method not specifically tailored to particular proteins, we first verified that the rule holds for our data set (Results).

Compiling charge-differences for best model. We implemented the positive-inside rule in the

following two steps. (1) All non-transmembrane (loop) regions for the best model resulting from the refinement procedure were grouped into odd and even loops. (2) The difference between the positive charges C in all even and all odd loop regions was compiled:

$$\Delta C = \sum_{l=1}^{\text{INT}(N_{\text{loop}}/2)} C_l - \sum_{l=0}^{\text{INT}(N_{\text{loop}}/2)} C_{l+1} \quad (8)$$

)

$$\text{with } C_l = \frac{1}{L_l} \sum_{k=1}^{L_l} \delta_k^{\text{R+K}} \delta_k^L,$$

$$\text{and } \delta_k^{\text{R+K}} = \begin{cases} 1, & \text{if loop residue } k \text{ either R or K} \\ 0, & \text{else} \end{cases}$$

where N_{loop} was the number of non-transmembrane regions, $\text{INT}(x)$ the integer value of x , and L_l the length of the l -th loop. For $\Delta C \leq 0$ the protein N-term (begin) was predicted to be extra-cytoplasmic, and for $\Delta C > 0$ to be intra-cytoplasmic.

Modifying topology prediction. For our final predictions, we added the following two corrections to eq. 8. (1) The sums over the charges C were compiled over the profiles obtained from the multiple sequence alignments, i.e., for all aligned sequences the percentages of positively charged residues were summed. (2) Globular regions of more than 60 residues obscure the positive-inside signal (von Heijne & Gavel 1988). Thus, only the first 15 (N-term) and the last 25 (C-term) residues of loops longer than 60 residues were considered when compiling the charge difference ΔC .

Defining the reliability of the topology prediction. In analogy to the reliability for the correctness of the best model (eq. 6), we defined the reliability of the topology prediction with respect to the second best model:

$$Ri_T = \Delta C^{\text{best model}} - \Delta C^{\text{second best model}} \quad (9)$$

where $\Delta C^{\text{best model}}$ was the difference between positively charged residues in even and odd loop regions for the best

Method	Set	Per-residue accuracy			Number of HTM's			Per-segment accuracy		Correctly predicted proteins	
	N_{prot}	Q_2	Cor_H	I	N_{obs}	N_{prd}	N_{cor}	$Q_H^{%obs}$	$Q_H^{%prd}$	Q_M	Q_T
PHDhtm no filter	68	92.3	0.78	0.58	257	226	196	76.3	86.7	41.2	39.7
PHDhtm filter	68	95.0	0.85	0.65	257	252	247	96.1	98.0	83.8	75.0
PHDhtm refined	68	93.9	0.80	0.57	257	265	253	98.4	94.4	86.8	83.8
Jones et al., 1994	83									79.5	77.1

Table 1. Accuracy for set of 68 proteins. The names of our 68 test proteins are listed in Table 2; those used by Jones et al. represent a super-set of the 68 (Jones et al., 1994). Abbreviations for methods: *PHDhtm no filter*, raw neural network results used as starting point for the refinement procedure; *PHDhtm filter*, neural network with empirical filter (Rost, et al. 1995); *PHDhtm refined*, refinement of neural network prediction as described here; *Jones et al., 1994*, prediction method based on statistics, results compiled from the original publication (Jones et al., 1994). Abbreviations for scores: N_{prot} , number of proteins in set; Q_2 , percentage of residues predicted correctly in either of the two states: HTM, or not-HTM; Cor_H , Matthews correlation coefficient for state H (Matthews 1975); I , information entropy of per-residue prediction

(Rost & Sander 1993)); N_{obs} , number of HTM's observed; $N_{S\ DO3}(prd)$, number of HTM's predicted; N_{cor} , number of HTM's correctly predicted; $Q_H^{%obs} = 100 * (\text{number of HTM's predicted correctly} / \text{number of HTM's observed})$, i.e., percentage of correctly predicted HTM's; $Q_H^{%prd} = 100 * (\text{number of HTM's predicted correctly} / \text{number of HTM's predicted})$, i.e., likelihood that predicted HTM's were predicted correctly; Q_M , percentage of proteins for which all HTM's were predicted correctly; Q_T , percentage of proteins for which all HTM's and the topology were predicted correctly. (Note: from left to right the scores capture increasingly more aspects about global structural aspects.)

model, and $\Delta C^{second\ best\ model}$ that for the second best model.

Results and discussion

Significant improvement by refining neural network output. The post-processing of the neural network output by the dynamic programming-like algorithm yielded a significant improvement over the simple winner-takes-all decision (prediction = output state with maximal value) made by PHDhtm. The per-residue accuracy was marginally better for the refinement procedure than for the PHDhtm (Table 1). However, the really significant leap was in the per-segment accuracy. For 86% of the test proteins the refined version of PHDhtm predicted all segments correctly, while the simple

network achieved a correct prediction of all HTM's only for 41% of the proteins (Table 1). In other words, the incorporation of global information by the refinement procedure more than doubled the percentage of proteins globally predicted correctly.

Better segment prediction by refinement than by filter. The refinement procedure achieved a similar goal as the empirical filter: too long helices were split. How did the performance of the refinement algorithm compare to the empirical filter, in detail? Per-residue accuracy of the filtered prediction was clearly higher than for the refined prediction (Table 1). However, the refinement was clearly superior to the filter in terms of the more global per-segment scores or the percentage of proteins for which all segments were correctly predicted (Table 1).

Second best model occasionally better than best model. One feature of the dynamic programming-like refinement was the generation of a list of possible models. For nine (of 68) proteins, the prediction of some HTM's was wrong. For two of these (both with one HTM: *il2b_human* and *myp0_human*, Table 2), the second best model was entirely correct. For another five the second best model was also incorrect but more accurate than the best model (*cyoa_ecoli*, *cyoe_ecoli*, *iggb_strsp*, *pt2m_ecoli*,

suis_human; Table 2). For six of the seven proteins for which the second best model was more accurate, the reliability of the predicted model (eq. 6) was below the average value of 0.044 (exception *myp0_human*; Table 2). In general, entirely correct and partly correct models separated in terms of the average reliability: for 59 correct models the average was 0.047, for nine partly false models the average was 0.021.

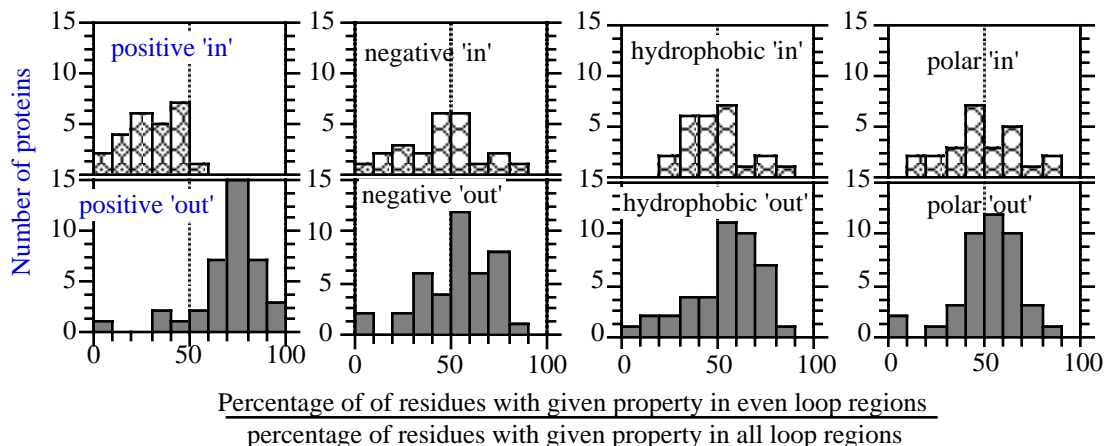


Fig. 4. Differences between intra- and extra-cytoplasmic regions. Distributions of the percentages of positive, negative, hydrophobic and polar residues shown separately for proteins with topology *in* and *out* (Fig. 1). The horizontal axes describe the excess of residues of a given bio-chemical property in even loop regions. Grey lines indicate the distinction between an abundance of a given residue type in intra- or in extra-cytoplasmic regions. For

example, most proteins with topology *in* have less positive residues in even loops than in odd loops (upper left quadrant; peak < 50%), whereas most proteins with topology *out* have more positive residues in even than in odd loops (upper right quadrant; peak > 50%). Hence, the differences in positive charges distinguish between intra- and extra-cytoplasmic regions.

Positive residues best discriminator for topology. For our data set of 68 proteins from various species, we could verify that positively charged residues occurred clearly more often in intra-cytoplasmic than in extra-cytoplasmic loops (Fig. 4). Other bio-chemical properties of amino acids (hydrophobicity, negative charge, polarity) could not be used successfully to distinguish between 'in' and 'out' (Fig. 4). Using the positive-inside rule and starting from the observed HTM's (i.e., an entirely correct

prediction) the topology was correctly assigned to 65 of the 68 proteins (false assignments occurred for *glra_rat*, *heam_measi*, and *lep_ecoli*).

Topology prediction unravels quality of refinement. Was it more important for the prediction of topology to predict residues more accurately (better per-residue accuracy of PHDhtm filter) or to predict the segments more accurately (better per-segment accuracy of PHDhtm refined)? The accuracy in predicting

topology was significantly better for the global refinement of the local neural network output than both for the simple network and for the filtered version (Table 1). (Note: given the HTM locations, a random prediction of topology would yield an accuracy of 52%.) The limiting factor was not the insufficiency of the positive-inside rule, but the incorrect prediction of the number of HTM's, as topology was predicted correctly for 97% of the proteins for which all HTM's had been correctly predicted.

Higher rate of false positives. The refinement method falsely predicted HTM's for more than 10% of proteins without HTM's (false positives). Thus, when using the method for an automatic prediction service (Rost 1996) we would still use the previously designed filter for the task of distinguishing proteins with and without HTM's (expected rate of false positives about 6% false positives; (Rost, et al. 1995). However, a similar idea as the one on which the refinement was based proved to enable the design of a method

tailored to reduce the rate of false positives significantly (Rost, et al. 1996).

Prediction accuracy compares favourably with other methods. One of the most accurate methods for predicting transmembrane helices and topology appears to be the one of Jones, Taylor and Thornton (Jones, et al. 1994). Since Jones et al. used a slightly different data set than we, a strict comparison with our method is not possible. However, the data sets overlap sufficiently to conclude that the refined post-processing of PHDhtm was more accurate than the method published by Jones et al. (Table 1).

Conclusion

Dynamic programming-like post-processing of neural network output successful. Post-processing the neural network output by a segment-oriented dynamic programming algorithm resulted in significantly better

name	N cor	N obs	N prd	Ri _M	Top obs	Top prd	Ri _T
1prc_H	1	1	1	0.056	out	out	11
1prc_L	5	5	5	0.054	in	in	-10
1prc_M	5	5	5	0.048	in	in	-7
4f2_human	1	1	1	0.030	in	out	2
5ht3_mouse	4	4	4	0.032	out	out	1
a1aa_human	7	7	7	0.001	out	out	16
a2aa_human	7	7	7	0.035	out	out	16
a4_human	1	1	1	0.021	out	out	3
aa1r_canfa	7	7	7	0.042	out	out	14
aa2a_canfa	7	7	7	0.000	out	out	16
adt_ricpr	12	12	12	0.012	in	in	-7
bach_halss	7	7	7	0.050	out	out	5
bacr_halha	7	7	7	0.059	out	out	6
cb21_pea	3	3	3	0.004	in	in	-1
cek2_chick	1	1	1	0.009	out	out	12
cyoa_ecoli	2	2	4	0.030	out	in	-6
cyob_ecoli	13	15	13	0.011	out	in	-3
cyoc_ecoli	5	5	5	0.075	in	in	-9
cyod_ecoli	3	3	3	0.107	in	in	-6
cyoe_ecoli	6	7	7	0.028	in	in	-3
edg1_human	7	7	7	0.040	out	out	12
egfr_human	1	1	1	0.008	out	out	13
fce2_human	1	1	1	0.046	in	in	-2
glp_pig	1	1	1	0.127	out	out	3
glpa_human	1	1	1	0.106	out	out	3
glpc_human	1	1	1	0.150	out	out	8
glra_rat	3	4	3	0.016	out	out	0
gmcr_human	1	1	1	0.040	out	out	8
gp1b_human	1	1	1	0.078	out	out	9
gpt_crilo	10	10	10	0.012	out	out	7
hema_cdvo	1	1	1	0.030	in	in	-5
hema_measi	1	1	1	0.029	in	in	-5
hema_pi4ha	1	1	1	0.025	in	in	-9
hg2a_human	1	1	1	0.035	in	in	-14

name	N cor	N obs	N prd	Ri _M	Top obs	Top prd	Ri _T
iggb_strsp	1	1	3	0.021	out	in	-14
il2a_human	1	1	1	0.059	out	out	34
il2b_human	1	1	2	0.016	out	in	-2
ita5_mouse	1	1	1	0.052	out	out	8
lacy_ecoli	12	12	12	0.006	in	in	-9
lech_human	1	1	1	0.048	in	in	-3
leci_mouse	1	1	1	0.044	in	in	0
lep_ecoli	2	2	2	0.045	out	out	13
mag1_mouse	1	1	1	0.025	out	out	3
malf_ecoli	8	8	8	0.020	in	in	-8
motb_ecoli	1	1	1	0.050	in	in	-11
mprd_human	1	1	1	0.072	out	out	0
myp0_human	1	1	2	0.056	out	in	-13
nep_human	1	1	1	0.022	in	in	-8
ngfr_human	1	1	1	0.032	out	out	16
oppb_salty	6	6	6	0.048	in	in	-12
oppc_salty	6	6	6	0.015	in	in	-13
ops1_calvi	7	7	7	0.031	out	out	10
ops2_drome	7	7	7	0.028	out	out	9
ops3_drome	7	7	7	0.029	out	out	9
ops4_drome	7	7	7	0.031	out	out	10
opsb_human	7	7	7	0.050	out	out	8
opsd_bovin	7	7	7	0.050	out	out	9
opsg_human	7	7	7	0.046	out	out	7
opsr_human	7	7	7	0.046	out	out	7
pigr_human	1	1	1	0.018	out	out	17
pt2m_ecoli	6	6	9	0.010	in	in	-4
sece_ecoli	3	3	3	0.095	in	in	-4
suis_human	1	1	3	0.003	in	in	-4
tcb1_rabbit	1	1	1	0.063	out	out	26
trbm_human	1	1	1	0.029	out	out	21

trsr_human	1	1	1	0.018	in	out	0
vmt2_iaann	1	1	1	0.200	out	out	8
vnb_inbbe	1	1	1	0.162	out	out	8

Table 2. Predicted and observed topology for 68 proteins. For the 68 transmembrane proteins used for cross-validation, the following data are listed: (1) protein name, given by the SWISS-PROT identifier (Bairoch & Boeckmann 1994); if the 3D structure is known, the PDB code plus chain identifier is used (Bernstein, et al. 1977,

Kabsch & Sander 1983); (2) number of transmembrane helices observed, predicted and correctly predicted; (3) reliability of refinement prediction (eq. 6); (4) observed topology (data taken from (Jones, et al. 1994)); (5) predicted topology; (6) reliability of topology prediction (eq. 9).

predictions of transmembrane helices than the original neural network. Furthermore, the refinement procedure solved the problem of too long helices predicted by the second level of neural networks (Rost, et al. 1995, Rost 1996) in a more eloquent manner than did the previously designed empirical filter (Rost, et al. 1995). The method involved only three free parameters (eq. 7), and was relatively stable with respect to the choice of these (data not shown).

Better per-segment prediction of refinement. Compared based on local per-residue scores the filtered version of PHDhtm was slightly superior to the refined version. However, the real strength of the refinement procedure was the successful use of global (in terms of sequence) constraints. The effect was an improvement of per-segment accuracy (Table 1). The superior performance in terms of HTM's made the refined PHDhtm more useful for the final topology prediction.

Better prediction of topology by refinement. Initially, we (i) trained neural networks to predict topology, and (ii) implemented a topology-prediction algorithm similar to the one described by Jones et al. (1994). Both methods were less successful than the simple positive-inside rule established by von Heijne (von Heijne & Gavel 1988). In combination with the prediction of transmembrane helices obtained by the refinement of PHDhtm, we succeeded to predict the topology correctly for more than 80% of the 68 test proteins (Table 1). False predictions of topology resulted primarily from false predictions of the number of helices: for 97% of the correctly predicted proteins the positive-inside rule yielded the correct topology.

Results only preliminary. The results presented here are preliminary in three respects. (1) The experimental assignment of transmembrane helices and topology is not

completely reliable (only four of the 68 test proteins are known by X-ray crystallography). (2) Although the refinement operated with very few free parameters, and although we chose those parameters on a subset of 10 of the 68 proteins, we still shall have to test the method on new proteins. (3) Performance accuracy would have to be compared to other (non-neural network) methods based on identical test sets.

Ready to digest entire genomes? In principle, the refined version of PHDhtm with the following topology prediction could be a valid tool for genome analysis. For experiments in molecular biology or pharmacology it may be important to know, e.g., all Haemophilus influenzae proteins that stick with both ends in the cell and have seven transmembrane helices. After completion of a more thorough evaluation of the refined version of PHDhtm and after reduction of the rate of false positives, the method described here will be ready to digest entire genomes (Rost, et al. 1996).

Method available. The refined version of PHDhtm and the topology prediction are available via an automatic prediction service (send the word *help* to the internet address *PredictProtein@EMBL-Heidelberg.DE*, or use the World Wide Web site (WWW) <http://www.embl-heidelberg.de/predictprotein/predictprotein.html>).

Generic results? We showed that a problem specific post-processing of the neural network output by dynamic programming resulted in more accurate predictions than did a simple winner-takes-all decision. This implied that the network output contained useful information about protein structure, that had been ignored by the simple network prediction. Would this hold up for other applications of neural networks, e.g., for the prediction of secondary structure and

solvent accessibility (Rost 1996)? The answer has yet to be investigated thoroughly. The algorithm proposed here was particularly simplified by the small number of segments in membrane proteins. Another interesting result was the answer to the question: which prediction was better 'refinement or filter'? Only when using global per-segment scores and when applying the method to predict topology, we could answer the question (refinement was superior). This sheds light on a general point: the evaluation of prediction method should mimic, as closely as possible, the demands of the user.

Acknowledgements. B.R. thanks Chris Sander (EMBL Heidelberg) for his financial support. Also, thanks to two of the referees for their valuable and detailed comments. Finally, thanks to all who deposit experimental results in public databases.

References

- Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W. and Lipman, D. J. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403-410.
- Bairoch, A. & Boeckmann, B. 1994. The SWISS-PROT protein sequence data bank: current status. *Nucl. Acids Res.* 22:3578-3580.
- Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T. and Tasumi, M. 1977. The Protein Data Bank: a computer based archival file for macromolecular structures. *J. Mol. Biol.* 112:535-542.
- Boyd, D. & Beckwith, J. 1990. The role of charged amino acids in the localization of secreted and membrane proteins. *Cell* 62:1031-1033.
- Casadio, R. & Fariselli, P. 1996. HTP: a neural network method for predicting the topology of helical transmembrane domains in proteins. *CABIOS* 12:41-48.
- Donnelly, D. & Findlay, J. B. C. 1995. Modelling alpha-helical integral membrane proteins. In Bohr, H. and Brunak, S. eds. *Protein folds: a distance based approach*. Boca Raton, Florida: CRC Press.
- Edelman, J. 1993. Quadratic minimization of predictors for protein secondary structure: application to transmembrane α -helices. *J. Mol. Biol.* 232:165-191.
- Engelman, D. M.; Steitz, T. A. and Goldman, A. 1986. Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annu. Rev. Biophys. Biophys. Chem.* 15:321-353.
- Fleischmann, R. D., et al. 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* 269:496-512.
- Hartmann, E.; Rapoport, T. A. and Lodish, H. F. 1989. Predicting the orientation of eukaryotic membrane-spanning proteins. *Proc. Natl. Acad. Sc. U.S.A.* 86:5786-5790.
- Hennessey, E. S. & Broome-Smith, J. K. 1993. Gene-fusion techniques for determining membrane-protein topology. *Curr. Opin. Str. Biol.* 3:524-531.
- Jones, D. T.; Taylor, W. R. and Thornton, J. M. 1994. A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochem.* 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.
- Kyte, J. & Doolittle, R. F. 1982. A simple method for displaying the hydrophathic character of a protein. *J. Mol. Biol.* 157:105-132.
- Lattman, E. E. 1994. Protein crystallography for all. *Proteins* 18:103-106.
- Manoil, C. & Beckwith, J. 1986. A genetic approach to analyzing membrane protein topology. *Science* 233:1403-1408.
- Matthews, B. W. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Ac.* 405:442-451.
- Persson, B. & Argos, P. 1994. Prediction of transmembrane segments in proteins utilising multiple sequence alignments. *J. Mol. Biol.* 237:182-192.
- Rost, B. 1995. Fitting 1-D predictions into 3-D structures. In Bohr, H. and Brunak, S. eds. *Protein folds: a distance based approach*. Boca Raton, Florida: CRC Press.
- Rost, B. 1996. PHD: predicting one-dimensional protein structure by profile based neural networks. *Meth. Enzymol.* 266:525-539.
- Rost, B.; Casadio, R. and Fariselli, P. 1996. Topology prediction for helical transmembrane proteins at 86% accuracy. *Prot. Sci.* submitted February, 1996.

Rost, B.; Casadio, R.; Fariselli, P. and Sander, C. 1995. Prediction of helical transmembrane helices at 95% accuracy. *Prot. Sci.* 4:521-533.

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. 1995. Progress of 1D protein structure prediction at last. *Proteins* 23:295-300.

Rost, B. & Sander, C. 1996. Bridging the protein sequence-structure gap by structure predictions. *Annu. Rev. Biophys. Biomol. Struct.* 25:in press.

Rost, B.; Sander, C. and Schneider, R. 1994. Redefining the goals of protein secondary structure prediction. *J. Mol. Biol.* 235:13-26.

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

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

Taylor, W. R.; Jones, D. T. and Green, N. M. 1994. A method for α -helical integral membrane protein fold prediction. *Proteins* 18:281-294.

von Heijne, G. 1986. The distribution of positively charged residues in bacterial inner membrane proteins correlates with the transmembrane topology. *EMBO J.* 5:3021-3027.

von Heijne, G. 1989. Control of topology and mode of assembly of a polytopic membrane protein by positively charged residues. *Nature* 341:456-458.

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.