Protein Science (1996), *5*:1704-1718. Cambridge University Press. Printed in the USA.

Topology prediction for helical transmembrane proteins at 86% accuracy

BURKHARD ROST,[1,2] PIERO FARISELLI,[3] AND RITA CASADIO[3]

[1] European Molecular Biology Laboratory, 69012 Heidelberg, Germany
[2] European Bioinformatics Institute, Hinxton Hall, Hinxton, Cambridge CB10 1RQ, England
[3] Department of Biology, Laboratory of Biophysics, University of Bologna, 40126 Bologna, Italy

(RECEIVED March 22, 1996; ACCEPTED May 31, 1996)

Abstract

Previously, we introduced a neural network system predicting locations of transmembrane helices (HTMs) based on evolutionary profiles (PHDhtm, Rost B, Casadio R, Fariselli P, Sander C, 1995, *Protein Sci 4*:521-533). Here, we describe an improvement and an extension of that system. The improvement is achieved by a dynamic programming-like algorithm that optimizes helices compatible with the neural network output. The extension is the prediction of topology (orientation of first loop region with respect to membrane) by applying to the refined prediction the observation that positively charged residues are more abundant in extra-cytoplasmic regions. Furthermore, we introduce a method to reduce the number of false positives, i.e., proteins falsely predicted with membrane helices. The evaluation of prediction accuracy is based on a cross-validation and a double-blind test set (in total 131 proteins). The final method appears to be more accurate than other methods published: (1) For almost 89% (±3%) of the test proteins, all HTMs are predicted correctly. (2) For more than 86% (±3%) of the proteins, topology is predicted correctly. (3) We define reliability indices that correlate with prediction accuracy: for one half of the proteins, segment accuracy raises to 98%; and for two-thirds, accuracy of topology prediction is 95%. (4) The rate of proteins for which HTMs are predicted falsely is below 2% (±1%). Finally, the method is applied to 1,616 sequences of *Haemophilus influenzae*. We predict 19% of the genome sequences to contain one or more HTMs. This appears to be lower than what we predicted previously for the yeast VIII chromosome (about 25%).

Keywords: dynamic programming; genome analysis; *Haemophilus influenzae*; postprocessing neural network output; secondary structure prediction; structure prediction for integral membrane proteins; topology prediction for helical transmembrane proteins

Integral membrane proteins comprise an important class of proteins for which experimental techniques for 3D structure determination are often not applicable. Fortunately, theoretical prediction of structural aspects is simpler for membrane proteins than for globular proteins because the lipid bilayer imposes strong constraints on the degrees of freedom for the 3D structure (von Heijne, 1981, 1989, 1992; Eisenberg et al., 1984; Engelman et al., 1986; von Heijne & Gavel, 1988; Taylor et al., 1994; Rost et al., 1995).

Reprint requests to: Burkhard Rost, EMBL, 69012 Heidelberg, Germany; e-mail: rost@embl-heidelberg.de.

Abbreviations: 1D, one-dimensional; 3D, three-dimensional; HTM, transmembrane helix (in figures and tables also abbreviated with the symbol H; L is used to describe nontransmembrane regions); PHDhtm, profile-based neural network prediction of helical transmembrane regions; PHDhtm_fil, empirical filter postprocessing the output from PHDhtm; PHDhtm_ref, refinement procedure postprocessing the output from PHDhtm described here.

Prediction of HTMs

3D structures are determined experimentally for two types of membrane proteins: (1) helical proteins consisting of typically apolar helices of about 20 residues that cross the membrane perpendicular to its surface [photo-reaction center (Deisenhofer et al., 1985); bacteriorhodopsin (Henderson et al., 1990); light harvesting complex II (Wang, 1994)], cytochrome *c* oxidase (Iwata et al., 1995); and (2) β proteins consisting of 16-stranded β-barrels [porin (Weiss & Schulz, 1992; Cowan & Rosenbusch, 1994; Kreusch & Schulz, 1994)]. Methods for the prediction of

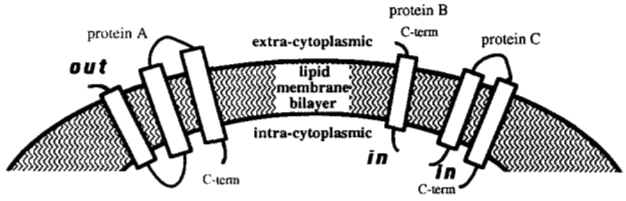

Fig. 1. Topology for helical transmembrane proteins. In one class of membrane proteins, typically apolar helical segments are embedded in the lipid bilayer oriented perpendicular to the surface of the membrane. The helices can be regarded as more or less rigid cylinders. The orientation of the helical axes, i.e., the topology of the transmembrane protein, can be defined by the orientation of the first N-terminal residues with respect to the cell. The topology is defined as *out* when the protein N-term starts on extracytoplasmic region (protein A) and as *in* if the N-term starts on the intracytoplasmic side (proteins B and C).

transmembrane segments usually focus on helical transmembrane proteins, for which more experimental data is available. Prediction methods were designed to predict the locations of HTMs (von Heijne, 1981, 1986a, 1986b, 1992; Argos et al., 1982; Kyte & Doolittle, 1982; Engelman et al., 1986; Cornette et al., 1987; von Heijne & Gavel, 1988; Degli Esposti et al., 1990; von Heijne & Manoil, 1990; Landolt-Marticorena et al., 1992; Donnelly et al., 1993; Edelman, 1993; O'Hara et al., 1993; Sipos & von Heijne, 1993; Jones et al., 1994; Persson & Argos, 1994; Donnelly & Findlay, 1995; Casadio et al., 1996) and the orientation of HTMs with respect to the cell (dubbed topology, Fig. 1; von Heijne & Gavel, 1988; von Heijne, 1989, 1992; Nilsson & von Heijne, 1990; Sipos & von Heijne, 1993; Jones et al., 1994; Casadio & Fariselli, 1996). If the locations of the HTMs and the topology are known with sufficient accuracy, 3D structure can be predicted successfully for the membrane spanning segments by an exhaustive search of the entire possible structure space (Taylor et al., 1994).

Accuracy of prediction methods

One of the problems in predicting structure for helical transmembrane proteins is the lack of accurate experimental information. Most prediction methods designed for globular water-soluble proteins are typically based on more than 100 proteins (Rost & Sander, 1994, 1995) of known 3D structure as stored in the Protein Data Bank (PDB) (Bernstein et al., 1977). To obtain sufficiently large data sets, prediction methods for membrane proteins use data from experimental sources other than crystallography or spectroscopy (Manoil & Beckwith, 1986; Park et al., 1992; Hennessey & Broome-Smith, 1993). There are numerous examples for proteins for which "reliable experimental information" obtained from different groups is contradictory. To list a few controversial cases: (1) nicotinic acetylcholine receptor channel: four α-helices versus two α-helices and two β-strands (Hucho et al., 1994); (2) P-type ATPases: 8 versus 10 α-helices (Stokes et al., 1994); (3) α-subunit of the FO channel *Escherichia coli*: topology *out* (Lewis et al., 1990) versus topology *in* (Bjorbaek et al., 1990); (4) mitochondrial cytochrome *b*: 7–9 α-helices (Degli Esposti et al., 1993). One consequence of this is that prediction methods are likely to become more accurate as reliable experimental information about integral membrane proteins is being added to the databases. Another consequence, however, is the problem to adequately estimate prediction accuracy. Thus, estimates for expected accuracy have to be taken with caution.

Are further improvements of prediction accuracy necessary?

Advanced methods for the prediction of HTMs (Jones et al., 1994; Persson & Argos, 1994; Rost et al., 1995) reach levels of about 90% accuracy (correctly predicted HTMs). Thus, predictions of HTMs are significantly more accurate than are two-state secondary structure predictions of, for example, helix, nonhelix for globular proteins (Rost & Sander, 1993b). Is there any need for further improvement of 1D predictions for transmembrane proteins? Indeed, two methods that start from 1D predictions of HTMs to predict further aspects of 3D structure would presumably benefit from better 1D predictions. (1) Taylor et al. (1994) achieve to predict 3D structure for the membrane spanning helices using the knowledge of the exact locations of the helices as the starting point. In general, current 1D predictions are not accurate enough to provide the demanded precision in locating the helices. (2) A simple and successful technique to predict topology is the positive-inside rule (von Heijne & Gavel, 1988; Hartmann et al., 1989; von Heijne, 1989, 1992; Boyd & Beckwith, 1990; Dalbey, 1990; Nilsson & von Heijne, 1990; Sipos & von Heijne, 1993): positively charged residues occur more often in intra-cytoplasmic than in extra-cytoplasmic regions. Applying this rule for the prediction of topology relies crucially on a correct prediction of the nontransmembrane regions. We shall show that relatively small improvements in 1D predictions of HTMs can result in significantly better predictions of topology.

An improvement and extension of a technique described previously to predict locations of HTMs (Rost et al., 1995) is presented here. The initial method (PHDhtm) used information derived from multiple sequence alignments as input for a system of neural networks (Fig. 2, step 1). The neural network preferences were used in two ways. (1) A region of 18 adjacent residues was searched that had the highest propensity in the protein to be in a transmembrane helix (Fig. 2, step 2). Then two thresholds were applied (Equation 5) to decide whether the protein was predicted to contain at least one HTM. (2) The preferences for HTM and not-HTM were input to a dynamic programming algorithm that produced a model (locations and number of HTMs) that was optimally compatible with the neural network preferences and the assumption that the protein contains HTMs of lengths 18–25 residues (Fig. 2, step 3; Figs. 6, 7). By working on the preferences for the entire protein, the refinement procedure introduced an aspect global in sequence, i.e., the resulting model was not as constrained to signals local in sequence (17 adjacent residues used as input to the neural networks) as the previous network prediction. Finally, the refinement model was used to predict topology (Fig. 1) by applying the positive-inside rule (Fig. 2, step 4; Fig. 6). The main elements of the method are described in mathematical details elsewhere (Rost et al., 1996). Here, we focused on the new aspects (reduction of false positives; definition of reliability indices for the prediction) and present a thorough analysis of the performance of the novel method. Finally, the tool was applied to the first entirely sequenced genome of *Haemophilus influenzae* (Fleischmann et al., 1995) and particular aspects of the results were com-

pared with an analysis of the yeast VIII chromosome (Rost et al., 1995).

Results and discussion

Correct prediction of all HTMs for almost 90% of the proteins

Refinement procedure significantly better than original neural network

The refinement algorithm (PHDhtm_ref) used here systematically optimized the transmembrane segments compatible with the output of the neural network system PHDhtm. The success was that the number of proteins for which all HTMs were predicted correctly almost doubled (Table 1). More than 98% of all observed HTMs were predicted correctly by PHDhtm_ref (337 of 341 observed; Table 1). Tendency was a marginal overprediction (341 observed, 354 predicted; Table 1). Prediction accuracy was higher for proteins that were observed to contain more than one HTM (data not shown).

Refinement procedure better at predicting segments than empirical filter

HTMs predicted by PHDhtm alone were too long (266 predicted versus 341 observed; Table 1). The reason is that loop regions between two transmembrane segments are often very hydrophobic. Because the neural network only "sees" biochemical properties of amino acids, the second level of neural networks introduced to account for correlations between adjacent residues (Rost et al., 1995; Rost, 1996a) frequently predicted helices extending to more than 40 residues. Thus, the network system could not learn external constraints imposed on the structure. Previously, we have corrected this shortcoming by introducing an empirical filter that simply chopped too-long helices into several shorter ones (Rost et al., 1995); PHDhtm_fil: 340 HTMs predicted versus 341 observed; Table 1). The refinement algorithm pursued systematically a similar goal. PHDhtm_ref predicted residues slightly less correctly than PHDhtm_fil, but was slightly better at correctly predicting HTMs (Table 1).

Expected accuracy verified by double-blind test

After we had completed all tests with the cross-validation set of 83 membrane proteins, we tested all methods on the double-blind set of 48 proteins. The results corrected our previous estimates for prediction accuracy to higher values. In particular, PHDhtm_ref performed even better when applied to a set of proteins that had never been used before (Table 1). (Note: Most results presented in the following hold for the entire set of 131 proteins, i.e., cross-validation plus double-blind set.)

Reliability index guide for expert-driven improvement of accuracy

The reliability index defined for the final best refined model (Equation 3) correlated well with prediction accuracy (Fig. 3). In practice, this allows focus on the subset of proteins that were predicted more reliably. For example, 66 proteins were predicted at levels of $Ri_M \geq 3$; for 65 of these 66 proteins, all predicted HTMs were correct (Fig. 3; outlier: myp0_human for which the signal peptide was predicted as HTM; see the Electronic Appendix or Rost, 1996b).

Table 1. *Accuracy of predicting transmembrane helices and topology*[a]

Method	Set N_{prot}	Number of transmembrane helices N_{obs}	N_{prd}	N_{cor}	Per residue accuracy Q_2	Per segment accuracy Q_M[b]	Accuracy for topology prediction Q_T[b]
PHDhtm_nof	83	341	300	266	91.9	**45.8** ± 6.0	**44.6** ± 6.0
PHDhtm_fil	83	341	340	333	94.5	**86.7** ± 3.6	**80.7** ± 4.8
PHDhtm_ref	83	341	354	337	93.6	**88.0** ± 3.6	**85.5** ± 4.8
Jones et al., 1994[c]	83					**79.5** ± 3.7	**77.1** ± 3.8
PHDhtm_fil	48	198	195	194	94.2	**89.6** ± 6.2	**85.4** ± 6.2
PHDhtm_ref	48	198	198	196	94.4	**91.7** ± 4.2	**87.5** ± 6.2
Eukaryotes[d]	99	334	337	332	95.8	**93.5** ± 3.2	**90.3** ± 3.2
Prokaryotes[d]	33	200	208	196	85.6	**75.8** ± 9.1	**72.7** ± 9.1
PHDhtm_fil	131	539	535	527	94.4	**88.5** ± 3.1	**82.4** ± 3.8
PHDhtm_ref	**131**	**539**	**552**	**533**	**93.8**	**89.3** ± **3.1**	**86.3** ± **3.1**

[a] Results given for cross-validation set [83 proteins; see the Electronic Appendix or Rost (1996b) #1152], double-blind set [48 proteins; see the Electronic Appendix or Rost (1996b) #1152], and for the sum of these two. Methods: PHDhtm_nof, neural network results (no filter); PHDhtm_fil, neural network with empirical filter (Rost et al., 1995); PHDhtm_ref, refined version of PHDhtm described here; Jones et al., 1994, prediction method of Jones et al. (1994). Scores and numbers: N_{prot}, number of proteins; N_{obs}, number of HTMs observed; N_{prd}, number of HTMs predicted; N_{cor}, number of HTMs correctly predicted; Q_2, percentage of residues predicted correctly in either of the two states, HTM or not-HTM; Q_M, percentage of proteins for which all HTMs were predicted correctly; Q_T, percentage of proteins for which the topology and all HTMs were predicted correctly. Note: As a rule of thumb, for an evaluation set of 131 proteins and 2 SDs of 2 × 3.1%, an improvement of >0.6% would be significant.

[b] Estimated error: ± x, where x was 1 SD for a binomial distribution.

[c] Results compiled from literature (Jones et al., 1994).

[d] Subsets with all eukaryotic and all prokaryotic proteins.

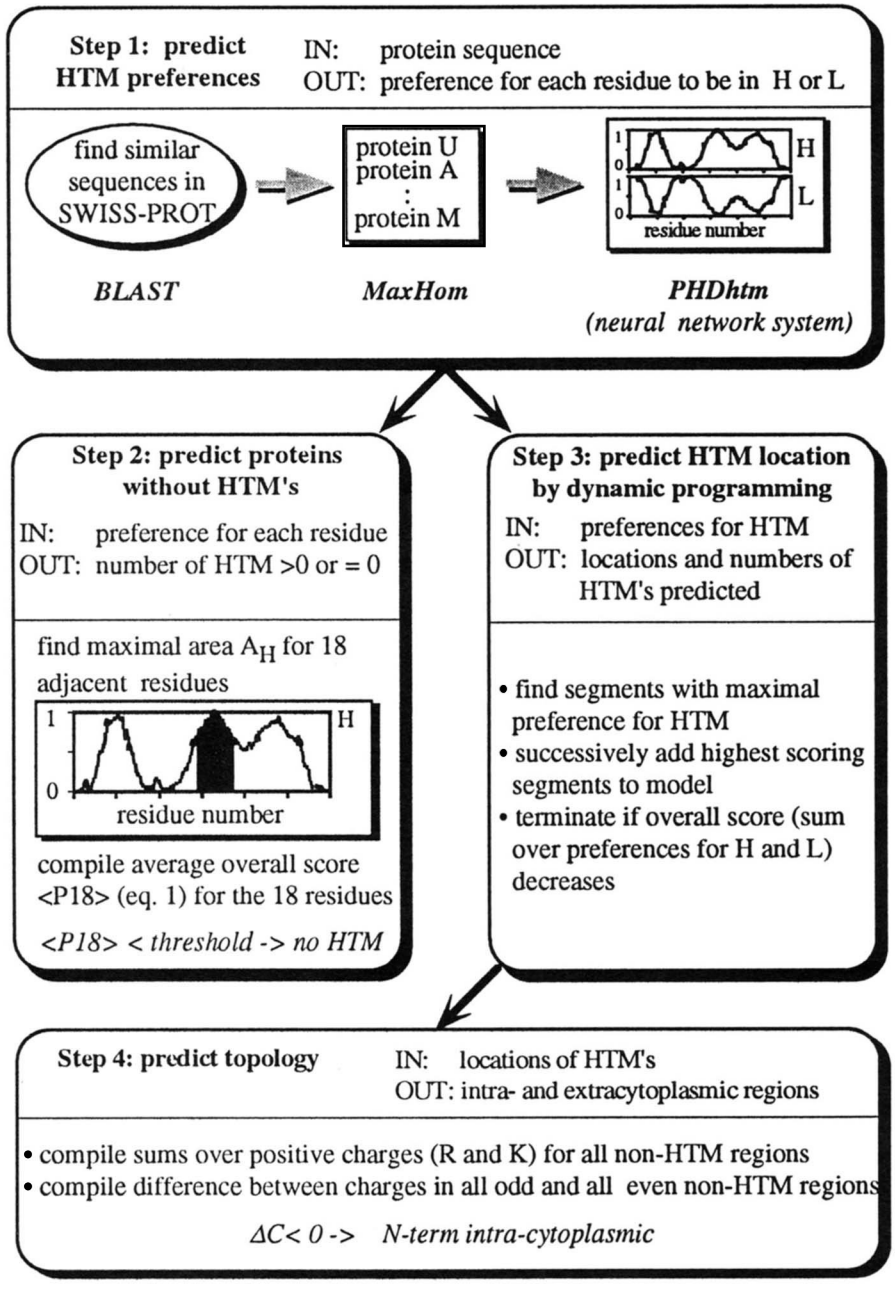

Fig. 2. From sequence to topology prediction. Step 1: Sequences similar to the input were found in SWISS-PROT (Bairoch & Boeckmann, 1994) using BLAST (Altschul et al., 1990; Karlin & Altschul, 1990); likely homologues were picked realigned by MAXHOM (Sander & Schneider, 1991, 1994) and the alignment was fed into the neural network system PHDhtm (Rost et al., 1995). The network preferences for each residue to be in a transmembrane helix (H) or to be outside of the lipid bilayer (L) were used as for the postprocessing methods described here. Step 2: The region of 18 adjacent residues with maximal preference for H was picked, normalized by the preferences for L, and a decision-threshold was applied to manage the distinction between proteins with and without HTMs. Step 3: The network preferences were used as input to a dynamic programming algorithm that found the model (number and locations of HTMs) representing the best path through all possible models consisting of HTMs between 18 and 25 residues by optimizing the compatibility of the model with the neural network outputs. Step 4: The final refined model output from the dynamic programming was used to apply the positive-inside rule (von Heijne & Gavel, 1988; von Heijne, 1992).

Second-best model occasionally correct

The dynamic programming-like algorithm yielded a list of possible models. Results reported refer to the best model (best according to Equation 1). However, the second-best model was occasionally better: 5 of the 14 proteins (of 131) predicted with errors (cox2_parde, ig1r_human, il2b_human, myp0_human, rfpb_salty; see the Electronic Appendix or Rost, 1996b) were predicted correctly by the second-best model. For another seven (of the 14), the second-best model was more accurate than the

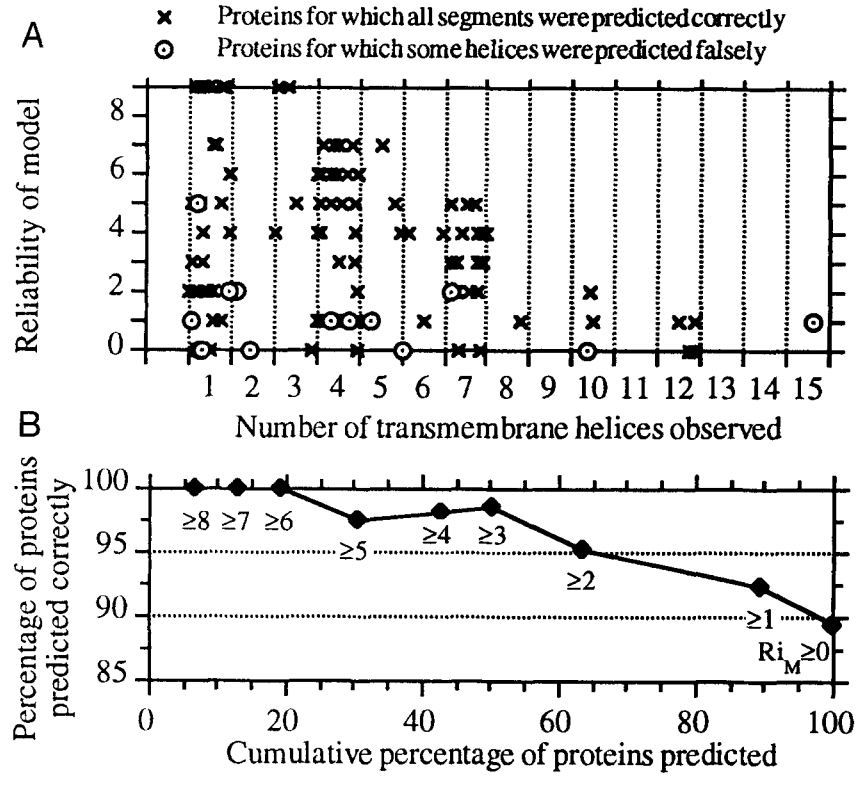

Fig. 3. Reliability of predicting correct model. **A:** Reliability of model versus number of HTMs observed. Note: To separate the points on the horizontal axis, we added a random number between 0 and 1 to the number of HTMs, i.e., all entries between two grey vertical lines represent the same number of helices. Crosses mark proteins for which all segments were predicted correctly; open circles proteins for which some helices were predicted falsely. For example, the highest index for a falsely predicted protein was 5 (myp0_human). **B:** Percentage of proteins for which all HTMs were predicted correctly versus the cumulative percentage of proteins predicted with a reliability index $Ri_M \geq n$, $n = 0$ (low), 1, . . . , 8 (high). $Ri_M \geq 0$ is the rightmost point representing 100% of the proteins. For example, more than 60% of all proteins were predicted with $Ri_M \geq 2$; for 95% of these, all HTMs were predicted correctly.

best. Thus, additional expert information may have had reduced the error from 11% to 7% or even to 2%. Expert decisions could have been based on the reliability index that was >2 for only 1 of the 14 proteins (myp0_human; for comparison: average reliability for all correctly predicted proteins = 3.4; Fig. 3; for details, see the Electronic Appendix or Rost, 1996b).

Correct topology prediction for more than 85% of the proteins

Refinement most successful in predicting topology

The empirical filter was slightly superior to the refinement in predicting residues, and slightly inferior in predicting segments. Which of these models (i.e., predictions of all HTMs) was more crucial for predicting topology? Using the refinement procedure as the basis for the positive-inside rule, we predicted topology correctly (and all HTMs) for 86% of all proteins (versus 82% for PHDhtm_fil; Table 1). Thus, PHDhtm_ref was significantly more useful as input for topology prediction than PHDhtm_fil. Furthermore, for more than 90% of the proteins the orientation of the first nonmembrane region was predicted correctly (data not shown; Note: A random prediction would be correct in about 52% of all cases).

Positive-inside rule not the limiting factor

For 117 proteins, all HTMs were predicted correctly; for 113 of these, the topology was predicted correctly. For three of the four proteins for which the predicted topology was not in accordance with the SWISS-PROT entries (4f2_human, Ih4_rhoac, and ssrg_rat), the application of the positive-inside rule yielded the wrong topology even when starting from HTM locations annotated in SWISS-PROT. The simple positive-inside rule yielded the correct topology for almost 97% of the proteins given HTM locations annotated in SWISS-PROT. Thus, the simplicity of the positive-inside rule was not the limiting factor for prediction accuracy.

Reliability index correlates with prediction accuracy

The value of the charge difference between extra- and intracytoplasmic nontransmembrane regions correlated with prediction accuracy (Fig. 4). The reliability index Ri_T (Equation 4) was >5 for only three falsely predicted proteins (myp0_human, iggb_strsp, and gaa4_bovin). For all three, some HTMs were predicted falsely (see the Electronic Appendix or Rost, 1996b). For only one of these three (myp0_human), the predicted model also had a high reliability, and thus could not have been suspected as a wrong prediction by an expert. For two

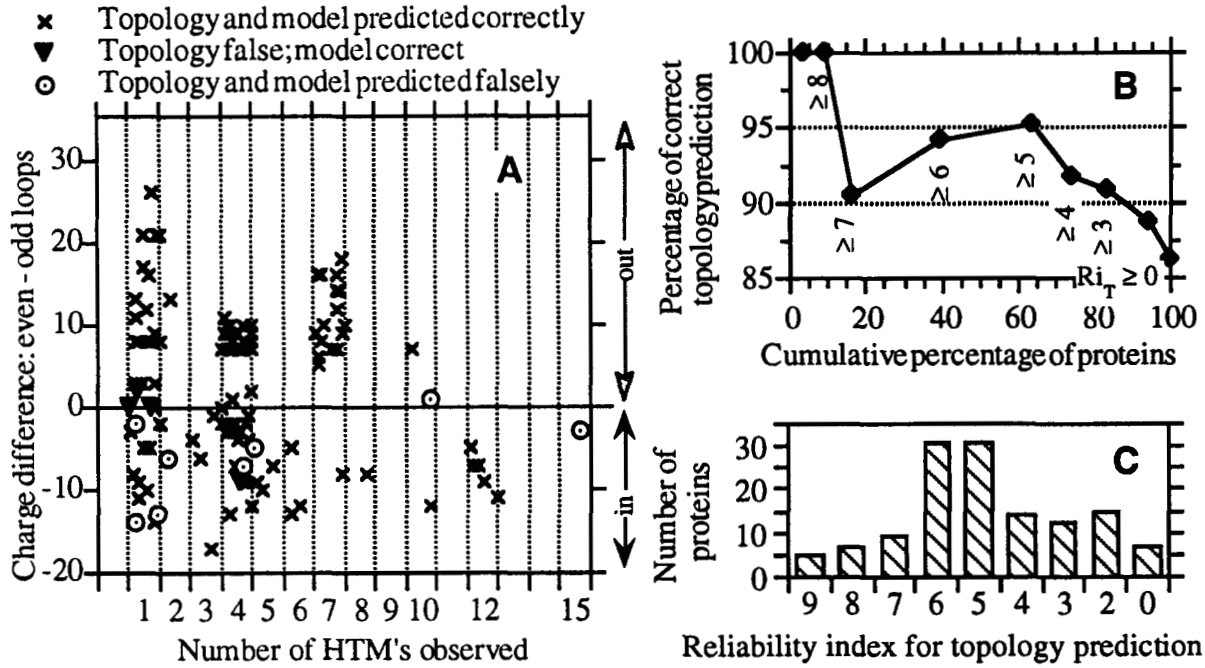

Fig. 4. Reliability of topology prediction. **A:** Charge difference versus number of HTMs observed. Note: To separate the points on the horizontal axis, we added a random number between 0 and 1 to the number of HTMs, i.e., all entries between two grey vertical lines represent the same number of helices. Crosses mark proteins for which all segments and the topology were predicted correctly; filled triangles mark proteins for which the topology prediction was wrong although all helices were correctly predicted; open circles mark proteins for which some helices and the topology were predicted falsely. High values for false topology predictions occurred only for proteins for which the model was also predicted falsely (circles). **B:** Accuracy of topology prediction versus the cumulative percentage of proteins predicted with a reliability index $Ri_T \geq n$, $n = 0$ (low), 1, . . . , 9 (high). $Ri_T \geq 0$ is the rightmost point representing 100% of the proteins. For example, more than 60% of all proteins were predicted with $Ri_T \geq 5$; for 95% of these, the topology and model were predicted correctly. **C:** The number of proteins predicted with a certain reliability index is shown to indicate that the drop of accuracy for $Ri_T \geq 6$ (B) is partly due to low count rates.

chains from the cytochrome *c* oxidase (cox1_parde and cox3_parde; see the Electronic Appendix or Rost, 1996b), we trusted our prediction more than the SWISS-PROT annotations for a homologue. The X-ray determination of the structure for cytochrome *c* oxidase (Iwata et al., 1995) revealed the correctness of the prediction (and consequently the mistake in SWISS-PROT; see details in the Electronic Appendix or Rost, 1996b).

Eukaryotic proteins predicted at higher accuracy

Separating the results for eukaryotic, prokaryotic, and viral proteins revealed three results. (1) Topology and all HTMs were predicted better than average for eukaryotic proteins (Table 1). (2) The positive-inside rule was about equally successful for both classes, i.e., given a correct prediction of all HTMs, the topology prediction was correct for 96.6% of the eukaryotic and for 96.0% of the prokaryotic proteins (Table 1). (3) The five viral proteins in our set were predicted correctly, although they all had single membrane spanning (expected accuracy below average; data not shown). However, five proteins are too few to justify any conclusion from this evidence. Why was prediction accuracy significantly higher for eukaryotes than for prokaryotes? We failed to find a satisfying answer. Several factors may have contributed to the higher accuracy for eukaryotes. (1) The multiple sequence alignments were more informative for the eukaryotes (20% of the alignments for eukaryotes had less than 4; 30% less than 10 sequences aligned; the respective numbers for prokaryotes: 40% and 70%!). (2) Eukaryotic HTMs are longer (on average 23 residues, versus 21 for prokaryotes; longer HTMs are predicted more reliably). (3) There are marginally more hydrophobic residues in eukaryotic HTMs (subclass of residues for which prediction accuracy was highest) and slightly more charged residues in eukaryotic non-HTM regions (second best predicted class of residues).

Reliable discrimination between proteins with and without HTMs

Significant reduction of false positives by evaluating strongest HTM

The usefulness of transmembrane predictions for the analysis of entire genomes depends crucially on the rate of false positives (i.e., proteins falsely predicted to contain HTMs). Here, we introduced a method tailored to reduce false positives. The method based on the hypothesis that proteins with and without HTMs separate most clearly when comparing a single region predicted with highest average propensity for HTM. Applying a strict decision threshold (Equation 5), the percentage of false positives was reduced below 2% (Table 2, note that the low rate of false positives was obtained at the expense of a higher false negative rate). False classifications occurred for proteins with very hydrophobic patches (for two of the falsely predicted seven

Table 2. *Accuracy of distinguishing proteins with and without transmembrane helices*[a]

Method	N_{glob}	E_{glob}[b]	N_{memb}	E_{memb}[b]
PHDhtm, $\vartheta^{strict} = 0.8$	435	**1.6% ± 0.7%**	131	**2.3% ± 1.5%**
PHDhtm, $\vartheta^{loose} = 0.7$	435	**3.7% ± 0.9%**	131	**0.0% ± 0.8%**
PHDhtm_fil	435	5.7% ± 1.1%	131	0.0% ± 0.8%
PHDhtm_fil[c]	278	4.3% ± 1.4%	69	0.0% ± 1.4%
Jones et al., 1994[c]	155	3.2% ± 1.9%	83	1.2% ± 1.2%
Edelman, 1993[c]	14	21.4% ± 14.3%		?

[a] Methods: PHDhtm, $\vartheta^{strict} = 0.8$: strict decision threshold applied to PHDhtm output (designed to reduce false positives; Equation 5); PHDhtm, $\vartheta^{loose} = 0.7$, loose decision threshold (designed to include all possible helical membrane proteins; Equation 5); PHDhtm_fil, PHDhtm plus empirical filter; Jones et al., 1994, statistics-based method for predicting HTMs (Jones et al., 1994); Edelman, 1993, statistics-based prediction method (Edelman, 1993); question mark indicates that published results for predicting membrane proteins are not based on cross-validation tests and thus are not comparable. Scores: N_{glob}, number of globular proteins, i.e., proteins without HTMs; E_{glob}, percentage of proteins without HTMs for which HTMs were predicted falsely; N_{memb}, number of proteins with HTMs; E_{memb}, percentage of proteins with HTMs for which no HTMs were predicted. The following proteins without HTMs were predicted to contain HTMs by the strict threshold: 1bmdA, oxidoreductase; 1pfiA, viral coat protein; 1ribA, reductase; 1spf, lipoprotein; 1ytbA, TATA-box binding protein; 2mnr, racemase; 2ohxA, oxidoreductase.

[b] Estimated error: ± x, where x was 1 SD for a binomial distribution.

[c] Results taken from literature (Edelman, 1993; Jones et al., 1994; Rost et al., 1995).

proteins, HTMs were predicted for observed strands: TATA-box binding protein, 1ytbA; and the racemase, 2mnr).

Will the estimate for false classifications hold for entire genomes?

The investigated set of 435 globular proteins resulted in more conservative estimates for the error rate than did smaller sets used previously (sets with 278, resp. 155 proteins; Table 2). The difference between the error rate for the maximal unique data set of 18 months ago and the maximal set used now (PHDhtm_fil for 238 versus 435 proteins, Table 2) indicates that the estimated rate of false positives should be viewed with skepticism. Improved experimental techniques may determine structures for proteins with very hydrophobic regions that could be predicted falsely as HTMs. Furthermore, the analysis is based on proteins contained in PDB that do not contain signal peptides, i.e., the problem that the refined prediction frequently confused HTMs and signal peptides is not taken into account. Thus, an expected rate of less than 2% false positives (Table 2) may prove to be too optimistic.

Total number of false classifications lower for strict threshold

The two decision thresholds introduced allow focus either on predicting as many helical transmembrane proteins as possible (loose threshold, Equation 5) or on minimizing the rate of false positives (strict threshold, Equation 5). The strict threshold was better in classifying proteins without HTMs (lower rate of false positives; higher rate of false negatives); the loose threshold in classifying proteins with HTMs (lower rate of false negatives; higher rate of false positives; Table 2). The strict threshold yielded a higher total error rate (false positives + false negatives = 3.9%) than the loose threshold (3.7%). However, for analyzing a large number of proteins by an automatic prediction service (Rost et al., 1994a; Rost, 1996a), (e.g., entire genomes) the total number of falsely classified proteins would be lower for the strict than for the loose threshold because the number of proteins without HTMs is supposedly below 30%.

Refined version of PHDhtm compared favorably with other methods

Better prediction of topology

The final topology predictions were more than eight percentage points superior to the best alternative method for prediction of topology published when evaluated on an identical data set of 83 proteins (Jones et al., 1994; Table 1). An empirically derived method was evaluated on 24 bacterial inner membrane proteins by von Heijne (1992). A crucial idea of that method was to choose the predicted HTMs such that the charge difference became maximal. In our hands, a similar algorithm resulted in significantly worse predictions than those obtained by the methods described here. The result published by von Heijne (1992) suggests a prediction accuracy of 96% for the correct prediction of all HTMs and topology. Omitting the three proteins for which the assignments of HTMs published by von Heijne did not correspond to the SWISS-PROT assignments (cyoa_ecoli, cyoe_ecoli, uhpt_ecoli), we achieved the same accuracy on this specially selected data set.

Lower rate of false positives

Judging from the results published, the method of Jones et al. (1994) is the best in distinguishing between proteins with and without HTMs. Our method tailored to manage this distinction yielded a lower error rate although based on a larger and more conservative data set (Table 2).

Analyzing the entire H. influenzae genome

Most predictions based on single sequence information

Prediction accuracy is significantly higher if the evolutionary information contained in multiple alignments is used as input to the neural network system PHDhtm (Rost et al., 1995). For 332 of 1,616 *H. influenzae* proteins, we predicted at least one HTM. For 129 of the 332 predicted HTM proteins (40%), the prediction was based on alignments; for only 76 (23%!), pre-

dictions were based on multiple alignments containing at least four sequences (results for the 37 of these predicted to contain at least two HTMs in Table 3; for more details, see the Electronic Appendix or Rost, 1996b). About 80% of predicted membrane-bound proteins (238) were predicted to contain more than a single HTM (see the Electronic Appendix or Rost, 1996c).

Fewer helical membrane proteins in H. influenzae than in yeast VIII

When subtracting the expected error rate for false positives (1.6 ± 0.7%; Table 2) and adding the expected underprediction of membrane proteins (2.3 ± 1.5%; Table 2), the results suggested that about 19% of all *H. influenzae* proteins contain HTMs; and about 16% more than one HTM. A similar analysis of the yeast VIII chromosome with our previous prediction method (PHDhtm_fil) predicted HTMs for about 25% of the proteins; and about 16% with more than one HTM. Given the higher error rate for false positives of our previous method (Table 2), the results suggested that there are slightly more proteins with HTMs in yeast VIII than in *H. influenzae*.

More proteins predicted with topology "in"

About 57% of the proteins predicted with HTMs were predicted with topology "in" (Fig. 5). Significant exceptions were

Table 3. *Proteins with transmembrane helices predicted for H. influenzae*[a]

Name	Top	N_{htm}	N-term	Segment positions
HI1586	out	13	MLSVLSINYR	29-52, 57-74, 87-104, 132-149, 154-177, 182-201, 216-238, 269-286, 311-330, 353-370, 390-408, 413-437, 493-511
HI0772	in	12	MISRVSRFMT	22-41, 56-76, 98-122, 136-153, 158-176, 191-212, 252-269, 274-298, 314-331, 336-360, 381-405, 417-441
HI0883	out	11	MTIESILSAI	14-35, 64-81, 86-106, 144-163, 180-203, 208-230, 235-259, 302-321, 348-370, 387-411, 416-433
HI1154	in	11	MLLVNLAIFI	29-47, 64-87, 102-126, 179-196, 219-243, 250-274, 284-301, 339-360, 365-389, 394-412, 417-434
HI0687	in	10	MNNENMVRVF	13-32, 37-57, 68-86, 93-116, 134-151, 163-180, 185-202, 228-245, 254-272, 277-294
HI1241	in	8	MSEQSSKYIA	12-33, 38-61, 72-96, 101-125, 130-147, 159-183, 188-206, 226-250
HI0359	out	7	MFDWLLEPLQ	19-39, 52-76, 96-113, 137-155, 174-198, 203-227, 235-259
HI0392	in	7	VDIFFVISGF	2-19, 35-59, 64-88, 95-119, 124-148, 161-182, 207-231
HI0407	out	7	MFEILFPALL	11-31, 42-66, 86-103, 128-148, 166-190, 195-218, 223-247
HI0825	out	7	MLINFTQVLQ	19-40, 61-84, 96-120, 131-155, 160-178, 183-201, 206-230
HI1248	in	7	MKKYKTGLVL	9-26, 56-76, 95-119, 135-153, 214-238, 250-270, 293-313
HI0188	in	6	MSNVDESQPL	24-42, 69-93, 110-134, 155-179, 190-207, 212-231
HI1122	in	6	MTDYRTQPIN	48-67, 108-132, 152-176, 181-198, 224-248, 278-297
HI1178	in	6	MFSDFLSLMF	15-36, 48-68, 86-104, 124-141, 146-165, 185-203
HI1187	in	6	MFKFVFKRIL	11-28, 99-120, 134-158, 200-218, 257-281, 302-323
HI1307	in	6	VMLNLIIVHL	1-23, 34-58, 63-86, 115-139, 144-168, 185-205
HI1548	in	6	MNTPFFISWR	28-52, 196-214, 270-292, 310-327, 332-354, 381-401
HI1621	in	6	MHLSEGVLHT	11-30, 35-59, 64-88, 93-117, 125-149, 163-187
HI1452	out	5	MEELLSAVII	26-49, 61-85, 90-108, 122-144, 162-186
HI1620	out	5	MKIHHLFQPH	8-27, 32-56, 61-85, 92-116, 121-145
HI0238	in	4	M19ISNYIH	15-39, 44-68, 73-92, 103-127
HI0318	in	4	MLFINITFAC	4-22, 33-54, 74-91, 130-150
HI0489	in	4	MDIFSFFSAD	14-38, 43-67, 91-109, 114-138
HI0976	out	4	MLYQILALLI	22-46, 60-82, 87-105, 110-128
HI1006	in	4	MSKKSGLSFL	9-26, 66-84, 95-113, 134-151
HI1602	in	4	MKDCKMQGIG	12-29, 47-71, 76-95, 112-131
HI0237	out	3	MLEMLKSWYS	22-41, 84-101, 157-177
HI0832	in	3	MVDQNPKRSG	23-43, 54-71, 94-111
HI0886	in	3	MNNLEKYRPY	17-34, 55-72, 98-116
HI1001	out	3	MDSRRSLLVL	347-366, 420-437, 496-515
HI1737	in	3	MTLIEQIITI	6-23, 41-58, 68-89
HI0484	out	2	METVITATII	12-32, 50-74
HI0633	in	2	MLWDLSGGMV	19-38, 43-63
HI1138	in	2	MKNKKLLVMA	103-121, 254-272
HI1594	out	2	MLIIGLCVVS	20-37, 42-66
HI1619	out	2	MMRCLFQAIG	17-34, 56-73

[a] We listed all proteins for which we predicted more than one HTM based on multiple sequence alignments [information for all 332 proteins predicted is in the Electronic Appendix or Rost (1996c) #1152]. Sequence names as in Fleischmann, et al. (1995); alignments from http://www.sander.embl-heidelberg.de/genequiz/haemophilus.htm l. N_{htm}, number of HTMs predicted; Top, predicted topology; N-term, first 10 residues of sequence; Segment positions, positions of predicted HTMs.

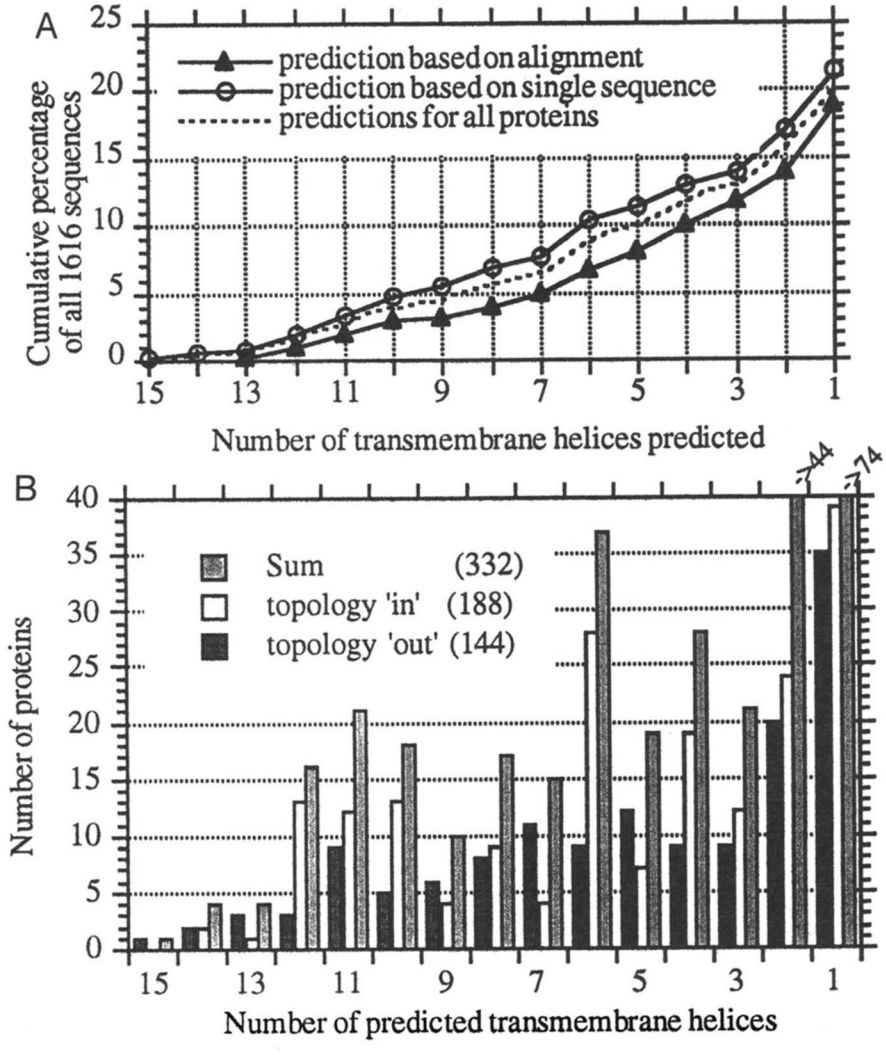

Fig. 5. Helical transmembrane proteins for *H. influenzae*. **A:** Cumulative percentage of helical transmembrane proteins versus number of HTMs predicted (total number of proteins 1,616). We separated between predictions based on multiple alignments (expected accuracy higher; filled diamonds) and predictions based on single sequence information only (expected accuracy lower; open diamonds; sums over all proteins given as dotted line). For example, 10% of the 1,616 *Haemophilus* sequences were predicted with ≥ 5 HTMs. **B:** Number of proteins predicted versus number of predicted HTMs. Open bars give all proteins predicted with topology *in*; filled grey bars all proteins predicted with topology *out*; dark bars give the sum over both.

proteins predicted with five and seven HTMs, for which the topology "out" dominated (Fig. 5). Interestingly, a higher percentage of the proteins predicted with topology "out" had both terminal nontransmembrane regions on the outside (86 of 144) than proteins predicted with topology "in" (79 of 188). In other words, proteins predicted with topology "out" were more often predicted with an odd number of HTMs.

Discussion

Significant improvement of prediction accuracy by refinement algorithm

The segment optimizing refinement of the profile-based neural network system PHDhtm proved to be successful in four ways. (1) Prediction accuracy was significantly better than for the simple neural network prediction; for about 89% (±3%, 1 SD) of the proteins, all HTMs were correctly predicted (Table 1). (2) The refined version of PHDhtm was significantly more accurate at predicting all HTMs correctly than was the previously implemented empirical filter (Table 1). (3) The refinement algorithm was less sensitive to the choice of free parameters (Equation 2) than the empirical filter because the results were better for the double-blind set that was used after the methods had been set up (Table 1). (4) The reliability index defined for the final prediction (Equation 3) correlated well with prediction accuracy: for 65 of the 66 (98%) proteins predicted with $Ri_M \geq 3$ all HTMs were predicted correctly (Fig. 3).

Prediction of topology better than 86% by combining refinement and positive-inside rule

The success of the refined version of PHDhtm showed most clearly for the prediction of topology (Fig. 1). (1) For more than

86% (±3%, 1 SD) of the proteins, all HTMs and the topology were predicted correctly (Table 1). (2) The limiting step for topology prediction was not the simplicity of the positive-inside rule: for 97% of the proteins for which all transmembrane regions had been predicted correctly, the positive-inside rule yielded the correct topology (Table 1). (3) The predicted reliability correlated well with accuracy: 83 proteins were predicted with a reliability ≥ 5 (Equation 4); for 79 of these the prediction was correct (Fig. 4). (4) Prediction accuracy was better than average for eukaryotic proteins (Table 1). (5) The final prediction of topology was significantly more accurate than the best alternative method published on a set of 83 eukaryotic and prokaryotic proteins (Jones et al., 1994). (6) A minor improvement in 1D accuracy resulted in a major improvement when using the 1D prediction to predict other aspects of protein structure. A similar effect, although less marked, is observed for prediction-based threading (Rost, 1995). One of the reasons for this effect is that the refinement algorithm successfully used information not local in sequence, i.e., extending over the windows of 17 adjacent residues input to the neural network system.

Reduction of false positives below 2% by evaluating strongest HTM

The analysis of entire genomes requires an accurate distinction between proteins with and without HTMs. Here we introduced an algorithm that distinguished the two classes based on a single helix for which PHDhtm predicted the highest average propensity (Equation 5). Less than 4% of the proteins were classified falsely by this procedure. In particular, for only 1.6% (±0.7%, 1 SD) from a large set of unique proteins (435) did we falsely predicted HTMs (false positives; Table 2). This was significantly better than our previous method and results published by others (Edelman, 1993; Jones et al., 1994). Lower rates of false positives implied higher rates of false negatives (proteins with HTMs that were not detected). The balance between the two can be shifted by switching between a strict threshold (1.6% false positives; 2.3% false negatives) and a loose threshold (3.7% false positives; 0% false negatives).

Method available by automatic prediction service

The refinement of PHDhtm and the topology prediction is available via an automatic prediction service (Rost et al., 1994a; Rost, 1996a); for information send the word "help" to the internet address PredictProtein@EMBL-Heidelberg.DE, or use the World Wide Web (WWW) site http://www.embl-heidelberg.de/predictprotein/. Alternative models are provided to enable expert users to focus on more reliably predicted HTMs. Note that it may lead to errors in predicting topology if the sequence starts or ends with HTM regions.

H. influenzae, an organism with few helical transmembrane proteins?

Finally, we scanned the entire *H. influenzae* genome (Fleischmann et al., 1995) for helical membrane proteins (CPU time for prediction: several hours on a SUN SPARC10). Given the error rate in distinguishing between proteins with and without HTMs (Table 2), the results suggested that about 19% of the *H. influenzae* proteins contain HTMs, and about 16% more than one HTM. These numbers were clearly lower than those obtained previously (Rost et al., 1995) for the entire yeast VIII chromosome (>25%). Will the difference in the percentage of helical membrane proteins between yeast and *H. influenzae* hold up for the entire genomes? What about the percentage of helical membrane proteins for other organisms? The tool to answer by dissecting genomes as they are being sequenced is set up.

Materials and methods

Database and evaluation of method

Selection of proteins

We based our analyses on proteins for which experimental information about the locations of HTMs is annotated in the SWISS-PROT database (Manoil & Beckwith, 1986; von Heijne & Gavel, 1988; von Heijne, 1992; Sipos & von Heijne, 1993; Bairoch & Boeckmann, 1994; Jones et al., 1994). The proteins were chosen to meet two criteria: (1) reliability: experimental information should be as reliable as possible (Manoil & Beckwith, 1986; von Heijne, 1992); and (2) comparability: the data set should be similar to those used by others (Jones et al., 1994). For the few known 3D structures, locations of HTMs were taken from DSSP (Kabsch & Sander, 1983). For all others, locations of HTMs are often controversial. In order to make the results easily reproducible for others, we decided to always use the definitions found in SWISS-PROT (Bairoch & Boeckmann, 1994). Locations and topology used are listed in the Electronic Appendix and on WWW (Rost, 1996b).

Cross-validation test

For the prediction of transmembrane propensities by the neural network system (PHDhtm, Rost et al., 1995), the cross-validation set of 83 transmembrane proteins was divided into 66 proteins used for training and 17 for evaluating the results (test set). This was repeated five times (fivefold cross-validation), until each protein had been in a test set once. The sets were separated such that no protein in the multiple alignments used for training had more than 25% pairwise sequence identity to any protein in the multiple alignments of the test proteins. The cross-validation procedure yields estimates for prediction accuracy that are likely to hold for proteins of yet unknown topology (Rost & Sander, 1993a, 1995).

Double-blind set

Although rigorous cross-validation experiments may yield sufficiently reliable estimates of prediction accuracy, prediction methods should always be evaluated additionally in a double-blind experiment, which proceeds in the following manner. First, the prediction method is developed and evaluated in a cross-validation experiment. Second, all parameters are frozen and the method is tested on a new set (double-blind set) of proteins that were not used before (ideally until the day the paper is submitted). We implemented this concept by: (1) optimizing free parameters on a subset of 10 proteins (chosen at random from cross-validation set); (2) compiling prediction accuracy by the cross-validation experiment; (3) evaluating the method on an additional double-blind set that was not used before. The double-blind set was selected by applying two criteria: (1) the entries for the locations of HTMs and topology should be labeled as "prob-

able" by the SWISS-PROT notation; and (2) for similar proteins of different species, only one protein was taken. The 48 proteins used as double-blind set are listed in the Electronic Appendix or Rost (1996b); all taken from SWISS-PROT release 32 (Bairoch & Boeckmann, 1994) that met those criteria and were not already contained in our cross-validation set.

Data for the H. influenzae genome

To illustrate the usefulness of our method, we report 332 "blind predictions" listing all proteins likely to contain HTMs for the entire *H. influenzae* genome. The sequences of the *H. influenzae* genome were taken from the TIGR Internet server (Fleischmann et al., 1995). The multiple sequence alignments for some of the 1,616 protein sequences of *H. influenzae* are publicly available (Casari et al., 1995).

Measuring prediction accuracy

In contrast to globular proteins for which the definition of segment-based scores for prediction accuracy is problematic (Rost et al., 1994b), evaluating methods predicting HTMs is relatively straightforward. Here we regarded an HTM to be predicted correctly if the overlap between observed and predicted helix was at least five residues.

Prediction methods

The dynamic programming-like algorithm and the prediction of topology are conceptually simple methods. Here we focused on describing the main idea of both methods and attempted to provide the details to the extent to make the work reproducible. A mathematically more explicit description is given elsewhere (Rost et al., 1996). The elements of the method introduced here were presented in more detail. These were the definitions of empirical reliability indices (1) for the prediction of the refined model, and (2) for the topology prediction; and (3) the new method to distinguish proteins with and without HTMs.

Neural network predictions of transmembrane preferences

Input for the refinement algorithm was the output of the profile-based neural network system PHDhtm (Fig. 2; Rost et al., 1995). The output of the networks consists of two values for each residue, giving the preferences of that residue to be in a transmembrane helix (H) or in a region outside of the lipid bilayer (L).

Finding the optimal path through all predicted propensities (dynamic programming)

The simplest way to derive predictions for helix locations from network preferences is to predict each residue to be in the state (H or L) with largest preference ("winner-takes-all" decision). The problem with this approach—that resulting HTMs were too long—was corrected by an empirical filter chopping too long helices into several shorter ones (Rost et al., 1995). A less arbitrary alternative for generating predictions from preferences is to find the optimal positioning of HTMs compatible with the network output (a similar dynamic programming method has been implemented for topology prediction by Jones et al., 1994). Because HTMs are observed to extend over about 18–25 residues, all possible HTMs can be enumerated. The dynamic programming-like algorithm was implemented by the following steps (Fig. 6).

1. Convert network output to propensity. The preferences (from PHDhtm) were normalized to propensities to yield preference H + preference L = 1 for each residue.
2. Compile pool of possible HTMs. The average propensity per helix was computed for all possible HTMs. Note that the number of possible helices is usually much larger than the number of residues (Fig. 7).
3. Generate models with increasing number of HTMs. Starting from the assumption that the protein contained no HTM ($\mu = 0$), we successively picked the best from the pool of all HTMs. Thus, models were generated with $\mu = 1, 2, \ldots, n$ HTMs (Fig. 7).
4. Select the best model. The final prediction was the model with highest sum over all propensities. The score P_μ for the model with μ helices was defined by:

$$P_\mu = \frac{1}{N_{res}} \sum_{k=1}^{N_{res}} p_k^H \delta_k^H + p_k^L \delta_k^L,$$

$$\text{with } \delta_k^L = 1 - \delta_k^H, \text{ and } \delta_k^H = \begin{cases} 1, \text{ if residue } k \text{ is in a helix,} \\ 0, \text{ else} \end{cases} \quad (1)$$

where N_{res} was the number of residues in the protein; p_k^H the propensities of residue k to be in a HTM, and p_k^L the propensity not to be in a HTM.

The algorithm is described based on three free parameters that were chosen by optimizing the performance of the method with respect to a subset of ten proteins. The parameters were the minimal and maximal length of HTMs, and the minimal length of a nontransmembrane region (dubbed loop) inserted between two helices. We used:

$$L^{min} = 18,\ L^{max} = 25,\ L^{loop} = 4. \quad (2)$$

Reliability index for best model

Instead of the reliability index associated with the network output for each residue (Rost, 1996a), here we introduced an index describing the reliability of the prediction for the correctness of the best model obtained by the refinement algorithm, i.e., the prediction that the protein has μ' helices. This index was based on the difference between the scores (Equation 1) for the best and for the second best model. We empirically favored the following definition:

$$Ri_M = INT(min\{9, 100 \times (P_{\mu'} - P_{\mu''})\}), \quad (3)$$

where $INT(x)$ was the integer value of variable x, $min\{x, y\}$ the minimum of x and y, $P_{\mu'}$ the score (Equation 1) for the best model predicting μ' HTMs, and $P_{\mu''}$ the score for the second best model predicting μ'' HTMs. Thus, the reliability adopted values between 0 (unreliable) and 9 (reliable).

Predicting topology based on the positive-inside rule

von Heijne established that membrane proteins of certain species contain more positively charged residues (arginine and lysine) on the intracytoplasmic side of the membrane than on the

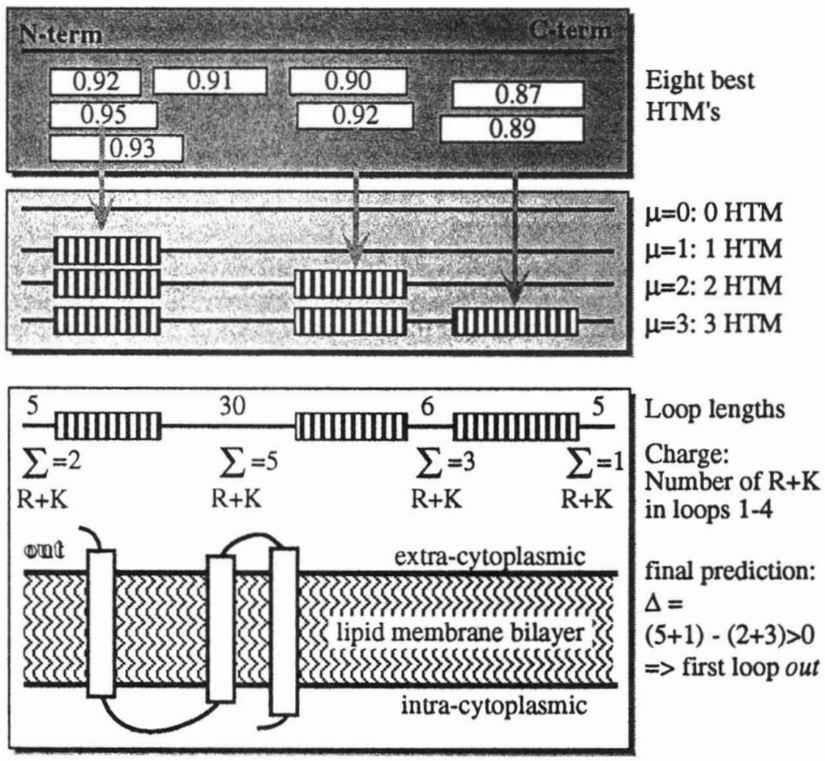

Fig. 6. Refinement of PHDhtm and prediction of topology. Refinement: the dynamic programming algorithm comprised the following three steps. (1) Compilation of pool of possible HTMs: for all possible HTMs, the preferences from the neural network output were summed over 18–25 residues; the results were stored (shown for best eight HTMs). (2) Generation of models with increasing number of HTMs: all possible models containing successively more helices, i.e., $\mu = 0, 1, \ldots, n$, were generated by selecting at each step μ that helix from the pool with maximal sum and no overlap to any of the helices added at previous steps $\mu' < \mu$. (3) Selection of best model: finally, the model μ with maximal sum over the network preferences was selected as prediction (here $\mu = 3$). Topology prediction: the number of positively charged residues (R, arginine; K, lysine) was summed separately over all odd (first, third, . . .) and over all even (second, fourth, . . .) nontransmembrane regions of the optimal model (highest sum over neural network preferences, here $\mu = 3$). The final prediction of topology was assigned according to the sign of the difference between the number of charged residues in odd and even regions. For example, for a positive difference, the first residues of the protein N-term were predicted as starting on the extracytoplasmic side.

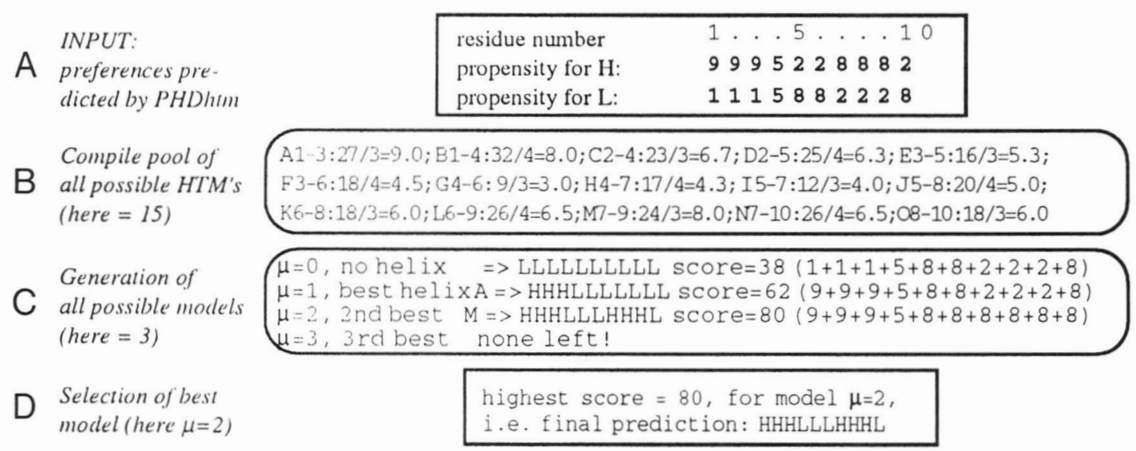

Fig. 7. Explicit example for the refinement algorithm. For simplicity, the following unrealistic parameters were used: minimal length of HTM = 3 residues; maximal length of HTM = 4 residues (for the real implementation, we used 18 residues for the minimal and 25 for the maximal length; Equation 2). **A:** Output from PHDhtm for a sequence of 10 residues converted to the propensities for each residue to be in a transmembrane helix (H) or not (L). **B:** Pool of all possible HTMs (A-O) of length 3 and 4; given are the numbers for the N- and C-term and the average helix propensity for each HTM. **C:** Starting from the model with no HTM ($\mu = 0$), successively the best HTMs are added; given the number of helices, the final prediction for all ten residues and the resulting score for that model (Equation 1). **D:** Best model is the one with $\mu = 2$ HTMs.

extracytoplasmic side (von Heijne & Gavel, 1988; von Heijne, 1989, 1992; Nilsson & von Heijne, 1990). Indeed, the rule was valid for more than 95% of the proteins in our data sets (data not shown). The application of this rule to the models obtained by PHDhtm (no filter), PHDhtm_fil, or the refined version of PHDhtm_ref required three steps (Fig. 6).

1. Compiling the positive charges. The positive charges C were compiled as percentages of positively charged residues (R and K) present in the entire sequence alignment of the protein. The percentages were summed separately for even and odd loop regions. (Note: For globular regions of more than 60 residues, we included only the 25 residues on the terminal sides.)
2. Computing the charge difference. The charge difference was compiled by subtracting positive charges of odd loop regions from positive charges of even regions (ΔC).
3. Predicting according to sign of charge difference. If the charge difference was negative ($\Delta C \leq 0$), the first loop was predicted to be extracytoplasmic; if it was positive ($\Delta C > 0$), to be intracytoplasmic.

Reliability index for predicting topology

The underlying hypothesis for defining a reliability index for the predicted topology was that the reliability would be proportional to the charge difference. We empirically favored the following definition:

$$Ri_T = INT\left(min\left\{9,\ 2 \times \sqrt{|\Delta C^2|}\right\}\right), \quad (4)$$

where $INT(x)$ was the integer value of x, $min\{x, y\}$ the minimum of x and y, and $|\Delta C|$ the absolute value of the charge difference. The definition normalizes the reliability index to values between 0 (unreliable) and 9 (reliable).

Distinguishing proteins with and without HTMs based on strongest HTM

Predictions of HTMs could be used to keep track of the flow of genome data (Oliver et al., 1992; Johnston et al., 1994; Fleischmann et al., 1995) by quickly scanning entire genomes for possible membrane associated proteins. For this purpose, we need methods to distinguish between proteins with and without HTMs. Previously, we used the empirical filter to accomplish the distinction (Rost et al., 1995; Rost, 1996a). The segment-oriented refinement algorithm provided an alternative solution to the problem that was applied to PHDhtm network output by the following three steps.

1. Converting output to propensities. For all residues, the neural network output was converted to propensities (i.e., preference H + preference L = 1).
2. Compiling propensity for best HTM. We scanned the protein for the segment of 18 (minimal length of HTM, Equation 2) consecutive residues with the maximal HTM propensity.
3. Applying decision thresholds. Finally, we predicted the protein to be globular if the average propensity for the best HTM was below a decision threshold ϑ.

We introduced two different thresholds for the decision to address two different possible goals of the user. (1) As many as possible helical membrane proteins should be found with as few as possible false positives (ϑ^{strict}). (2) All helical membrane proteins should be found even at the expense of including many false positives in the list (ϑ^{loose}). The following values were used:

$$\vartheta^{strict} = 0.8, \text{ and } \vartheta^{loose} = 0.7. \quad (5)$$

Results will be given for both constants (Table 2).

Supplementary material in Electronic Appendix

Folder name: rost-suppl.folder. File type: ASCII. Content: rost-setCross.txt, cross-validation set of 83 proteins, observed and predicted HTM locations and topology; rost-setBlind.txt, double-blind set of 48 proteins, observed and predicted HTM locations and topology; rost-hiAli.txt, predictions of HTM location and topology for all 129 *H. influenzae* proteins for which HTMs were predicted based on alignments. Note: In total, for 332 of the 1,616 *H. influenzae* proteins, HTMs were predicted; rost-hiNoAli.txt, predictions of HTM location and topology for all 203 *H. influenzae* proteins for which HTMs were predicted based on single-sequences. Note: In total, for 332 of the 1,616 *H. influenzae* proteins, HTMs were predicted.

Acknowledgments

B.R. thanks Chris Sander (EBI Cambridge) for financial support; Reinhard Schneider (EMBL Heidelberg) for providing the latest version of the alignment program MaxHom; Antoine de Daruvar (EMBL Heidelberg) for his assistance in running the PredictProtein service; and the GeneQuiz consortium (Heidelberg-Madrid-Menlo Park-Cambridge), especially Georg Casari and Reinhard Schneider (EMBL Heidelberg), for the *H. influenzae* sequence alignments. We thank all those who deposit experimental results in public databases and who maintain these databases, in particular, thanks to Amos Bairoch (Basel) and colleagues: the results of this work crucially depended on the quality of SWISS-PROT.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol 215*:403–410.

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-Protein protein sequence data bank: Current status. *Nucleic Acids Res 22*:3578–3580.

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.

Bjorbaek C, Foersom V, Michelsen O. 1990. The transmembrane topology of the alpha subunit from ATPase in *Escherichia coli* analyzed by PhoA protein fusions. *FEBS Lett 260*:31–34.

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. *Comput Appl Biosci 12*:41–48.

Casadio R, Fariselli P, Taroni C, Compiani M. 1996. A predictor of transmembrane α-helix domains of proteins based on neural networks. *Eur J Biophys 24*:165–178.

Casari G, Andrade MA, Bork P, Boyle J, Daruvar A, Ouzounis C, Schneider R, Tamames J, Valencia A, Sander C. 1995. Challenging times for bioinformatics. *Nature 376*:647–648.

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.

Dalbey RE. 1990. Positively charged residues are important determinants of membrane protein topology. *Trends Biochem Sci 15*:253–257.

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

Degli Esposti M, De Vires S, Crimi M, Ghelli A, Patarnello T, Meyer A. 1993. Mitochondrial cytochrome *b*: Evolution and structure of the protein. *Biochim Biophys Acta 1143*:243–271.

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.

Donnelly D, Findlay JBC. 1995. Modelling alpha-helical integral membrane proteins. In: Bohr H, Brunak S, eds. *Protein folds: A distance based approach*. Boca Raton, Florida: CRC Press. pp 155–164.

Donnelly D, Overington JP, Ruffle SV, Nugent JHA, Blundell TL. 1993. Modeling α-helical transmembrane domains: The calculation and use of substitution tables for lipid-facing residues. *Protein Sci 2*:55–70.

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. 1984. Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J Mol Biol 179*:125–142.

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

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, McKenney K, Sutton G, FitzHugh W, Fields C, Gocayne JD, Scott J, Shirley R, Liu LI, Glodek A, Kelley JM, Weidman JF, Phillips CA, Spriggs T, Hedblom E, Cotton mD, Utterback TR, Hanna MC, Nguyen DT, Saudek DM, Brandon RC, Fine LD, Fritchman JL, Fuhrmann JL, Geoghagen NSM, Gnehm CL, McDonald LA, Small KV, Fraser CM, Smith HO, Venter JC. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science 269*:496–512.

Hartmann E, Rapoport TA, Lodish HF. 1989. Predicting the orientation of eukaryotic membrane-spanning proteins. *Proc Natl Acad Sc USA 86*:5786–5790.

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.

Hennessey ES, Broome-Smith JK. 1993. Gene-fusion techniques for determining membrane-protein topology. *Curr Opin Struct Biol 3*:524–531.

Hucho F, Goerne-Tschelnokow U, Strecker A. 1994. Beta-structure in the membrane spanning part of the nicotinic acetylcholine receptor. *Trends Biochem Sci 19*:383–387.

Iwata S, Ostermeier C, Ludwig B, Michel H. 1995. Structure at 2.8 Å resolution of cytochrome *c* oxidase from paracoccus denitrificans. *Nature 376*:660–669.

Johnston M, Andrews S, Brinkman R, Cooper J, Ding H, Dover J, Du Z, Favello A, Fulton L, Gattung S, Geisel C, Kirsten J, Kucaba T, Hillier L, Jier M, Johnston L, Langston Y, Latreille P, Louis EJ, Macri C, Mardis E, Menezes S, Mouser L, Nhan M, Rifkin L, Riles L, Peter HS, Trevaskis E, Vaughan K, Vignati D, Wilcox L, Wohldman P, Waterston R, Wilson R, Vaudin M. 1994. Complete nucleotide sequence of *Saccaromyces cerevisiae* chromosome VIII. *Science 265*:2077–2082.

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.

Karlin S, Altschul SF. 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sc USA 87*:2264–2268.

Kreusch A, Schulz GE. 1994. Refined structure of the porin from *Rhodopseudomonas blastica*. *J Mol Biol 243*:891–905.

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.

Lewis MJ, Chang JA, Simoni RD. 1990. A topological analysis of subunit A from *Escherichia coli* F1F0-ATP synthase predicts eight transmembrane segments. *J Biol Chem 265*:10541–10550.

Manoil C, Beckwith J. 1986. A genetic approach to analyzing membrane protein topology. *Science 233*:1403–1408.

Nilsson IM, von Heijne G. 1990. Fine-tuning the topology of a polytopic membrane protein. Role of positively and negatively charged residues. *Cell 62*:1135–1141.

O'Hara PJ, Sheppard PO, Thøgersen 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, van der Aart QJM, Agostioni-Carbone ML, Aigle M, Alberghina L, Alexandraki D, Antoine G, Anwar R, Ballesta JPG, Benit P, Berben G, Bergantino E, Biteau N, Bolle PA, Bolotin-Fukuhara M, Brown A, Brown AJP, Buhler JM, Carcano C, Carignani G, Cederberg H, Chanet R, Contreras R, Crouzet M, Daignan-Fornier B, Defoor E, Delgado M, Demolder J, Doira C, Dubois E, Dujon B, Dusterhoft A, Erdmann D, Esteban M, Fabre F, Fairhead C, Faye G, Feldmann H, Fiers W, Francingues-Gaillard MC, Franco L, Frantali L, Fukuhara H, Fuller LJ, Galland P, Gent ME, Gigot D, Gilliquet V, Glansdorff N, Goffeau A, Grenson M, Grisanti P, Grivell LA, Haan Md, Haasemann M, Hatat D, Hoenicka J, Hegeemann J, Herbert CJ, Hilger F, Hohmann S, Hollenberg CP, Huse K, Iborra F, Indge KJ, Isono K, Jacq C, Jacquet M, James CM, Jauniaux JC, Jia Y, Jimenez A, Kelly A, Kleinhans U, Kreisl P, Lanfranchi G, Lewis C, van der Linden CG, Lucchini G, Lutzenkirchen K, Maat MJ, Mallet L, Mannhaupt G, Martegani E, Mathieu A, Maurer CTC, McDonnell D, McKee RA, Messenguy F, Mewes HW, Molemans F, Montague MA, Muzi Falconi M, Navas L, Newlon CS, Noone D, Pallier C, Panzeri L, Pearson BM, Perea J, Philippsen P, Pierard A, Planta RJ, Plevani P, Poetsch B, Pohl F, Purnelle B, Ramezani Rad M, Rasmussen SW, Raynal A, Remacha M, Richertich P, Roberts AB, Rodriquez F, Sanz E, Schaaff-Gerstenschlager I, Scherens B, Schweitzer B, Shu Y, Skala J, Slonimski PP, Sor F, Soustelle C, Spiegelberg R, Stateva LI, Steensma HY, Steiner S, Thierry A, Thireos G, Tzermia M, Urrestarazu LA, Valle G, Vetter I, van Vliet-Reedijk JC, Voet M, Volckaert G, Vreken P, Wang H, Warmington JR, Wettstein Dv, Wicksteed BL, Wilson C, Wurst H, Xu G, Yoshikawa A, Zimmermann FK, Sgouros JG. 1992. The complete DNA sequence of yeast chromosome III. *Nature 357*:38–46.

Park K, Perczel A, Fasman GD. 1992. Differentiation between transmembrane helices and peripheral helices by the deconvolution of circular dichroism spectra of membrane proteins. *Protein Sci 1*:1032–1049.

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

Rost B. 1995. TOPITS: Threading one-dimensional predictions into three-dimensional structures. In: Rawlings C, et al., eds. *Third International Conference on Intelligent Systems for Molecular Biology*. Cambridge, England/Menlo Park, California: AAAI Press. pp 314–321.

Rost B. 1996a. PHD: Predicting one-dimensional protein structure by profile based neural networks. *Methods Enzymol 266*:525–539.

Rost B. 1996b. *Appendix to "Topology prediction for helical transmembrane proteins at 86% accuracy."* EMBL, WWW document (http://www.embl-heidelberg.de/~rost/pap/ProtSci-05-96.html).

Rost B. 1996c. *Prediction of topology for all helical transmembrane proteins of Haemophilus influenzae*. EMBL, WWW document (http://www.embl-heidelberg.de/~rost/hi.html).

Rost B, Casadio R, Fariselli P. 1996. Refining neural network predictions for helical transmembrane proteins by dynamic programming. In: Rawlings C, et al., eds. *Fourth International Conference on Intelligent Systems for Molecular Biology*. St. Louis, Missouri/Menlo Park, California: AAAI Press. [submitted Jan 30, 1996].

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

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

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

Rost B, Sander C. 1994. Structure prediction of proteins – Where are we now? *Curr Opin Biotechnol 5*:372–380.

Rost B, Sander C. 1995. Progress of 1D protein structure prediction at last. *Proteins Struct Funct Genet 23*:295–300.

Rost B, Sander C, Schneider R. 1994a. PHD – An automatic server for protein secondary structure prediction. *Comput Appl Biosci 10*:53–60.

Rost B, Sander C, Schneider R. 1994b. 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 structurally meaning of sequence alignment. *Proteins Struct Funct Genet 9*:56–68.

Sander C, Schneider R. 1994. The HSSP database of protein structure-sequence alignment. *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.

Stokes DL, Taylor WR, Green NM. 1994. Structure, transmembrane topology and helix packing of P-type ion pumps. *FEBS Lett 346*:32–38.

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. 1986a. The distribution of positively charged residues in bacterial inner membrane proteins correlates with the trans-membrane topology. *EMBO J 5*:3021–3027.
von Heijne G. 1986b. A new method for predicting signal sequence cleavage sites. *Nucleic Acids Res 14*:4683–4690.
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.
von Heijne G, Manoil C. 1990. Membrane proteins—From sequence to structure. *Protein Eng 4*:109–112.
Wang ZX. 1994. Assessing the accuracy of protein secondary structure. *Nature Struct Biol 1*:145–146.
Weiss MS, Schulz GE. 1992. Structure of porin refined at 1.8 Å resolution. *J Mol Biol 227*:493–509.