

---

# Long membrane helices and short loops predicted less accurately

---

CHIEN PETER CHEN<sup>1</sup> AND BURKHARD ROST<sup>1,2,3</sup>

<sup>1</sup>CUBIC, Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY 10032, USA

<sup>2</sup>Columbia University Center for Computational Biology and Bioinformatics (C2B2), Russ Berrie Pavilion, New York, NY 10032, USA

<sup>3</sup>North East Structural Genomics Consortium (NESG), Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY 10032, USA

(RECEIVED May 5, 2002; FINAL REVISION September 16, 2002; ACCEPTED September 16, 2002)

## Abstract

Low-resolution experiments suggest that most membrane helices span over 17–25 residues and that most loops between two helices are longer than 15 residues. Both constraints have been used explicitly in the development of prediction methods. Here, we compared the largest possible sequence—unique data sets from high- and low-resolution experiments. For the high-resolution data, we found that only half of the helices fall into the expected length interval and that half of the loops were shorter than 10 residues. We compared the accuracy of detecting short loops and long helices for 28 advanced and simple prediction methods: All methods predicted short loops less accurately than longer ones. In particular, loops shorter than 7 residues appeared to be very difficult to detect by current methods. Similarly, all methods tended to be more accurate for longer than for shorter helices. However, helices with more than 32 residues were predicted less accurately than all other helices. Our findings may suggest particular strategies for improving predictions of membrane helices.

**Keywords:** Membrane proteins; protein structure prediction; predicting transmembrane helices; bioinformatics

*Predictions of membrane helices relatively successful.* Despite the great biological and medical importance of helical membrane proteins, we still know few three-dimensional (3D) structures. Fortunately, bioinformatics can contribute substantially to bridging the gap between what we do and

what we want to know by predicting membrane helices. In fact, predicting the locations of transmembrane helices (TMH) appears to be a simpler problem than predicting globular helices (Rost 1996, 2001). Nevertheless, although some investigators estimated the levels of accuracy to reach an incredibly high value of 99% (Jayasinghe et al. 2001), recent re-evaluations of many prediction methods (Ikeda et al. 2001; Möller et al. 2001; Chen et al. 2002) somewhat dampened this optimism by concluding that only the very best advanced methods predict all membrane helices correctly for >70% of all proteins, and that simple hydrophobicity scale-based methods tend to be ~20 percentage points less accurate.

*Distribution of membrane helix length crucial parameter for prediction.* Prediction methods typically explore that TMH are predominantly apolar and believed to be between 17 and 25 residues long (von Heijne 1996). The upper and lower bounds for the length of membrane helices are explicitly used by most prediction methods in two ways. (1)

---

Reprint requests to: Burkhard Rost, Columbia University, New York, NY 10032, USA; e-mail: rost@columbia.edu; fax: (212) 305-7932.

*Abbreviations:* 3D, three-dimensional; DSSP, program assigning secondary structure (Kabsch and Sander 1983); HMM, hidden Markov model; PDB, Protein Data Bank of experimentally determined 3D structures of proteins (Bernstein et al. 1977; Berman et al. 2000); SWISS-PROT, database of protein sequences (Bairoch and Apweiler 2000); TM, transmembrane; TMH, transmembrane helix.

*Terminology:* Advanced prediction methods: all methods that do not exclusively use a hydrophobicity scale; simple prediction methods: membrane prediction methods exclusively based on hydrophobicity scales; loop: referring to the region that connects two transmembrane helices in sequence; in particular, such loops could consist of entire structural domains.

Article and publication are at <http://www.proteinscience.org/cgi/doi/10.1110/ps.0214602>.

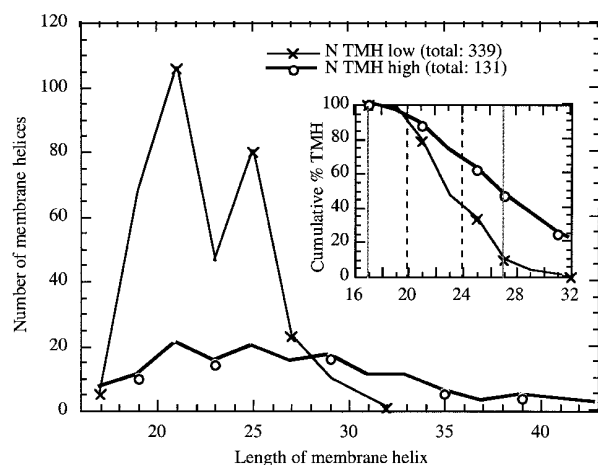
Some methods identify only hydrophobic regions as membrane helices that fall into the typical length interval (von Heijne 1992; Casadio et al. 1996; Persson and Argos 1996; Hirokawa et al. 1998; Ikeda et al. 2001; Jayasinghe et al. 2001). (2) Other methods search the best path through some predicted membrane helix propensity landscape that is compatible with such upper and lower bounds (Jones et al. 1994; Rost et al. 1996a,b; Krogh et al. 2001; Tusnady and Simon 2001). James Bowie found that the length distribution of three high-resolution structures was shifted toward longer helices (Bowie 1997).

Here, we re-evaluated the distribution of the length of TMH and that of the loops in between helices based on significantly larger dataset than previously used (Bowie 1997). Then, we analyzed 28 prediction methods in terms of their performance on short loops and long membrane helices.

## Results and Discussion

### *Many long helices and short loops observed in high-resolution structures*

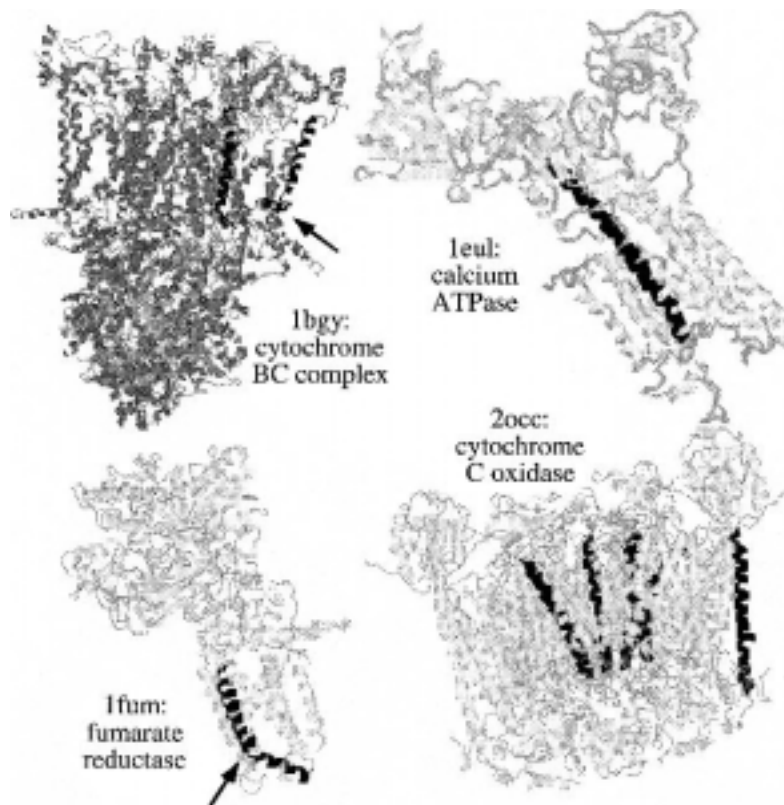
**Many membrane helices longer than 32 residues!** Half of all membrane helices annotated by low-resolution experiments were 20–24 residues long, whereas only about one-fourth of the high-resolution helices fall into this length interval (Fig.



**Fig. 1.** Length distributions for membrane helices. The lengths of the membrane helices were assigned using DSSP for the high-resolution data (36 unique chains; 131 helices), and using the annotation in SWISS-PROT for the low-resolution data (165 unique proteins; 339 helices). The inset gives the cumulative percentages of helices. About half the high-resolution helices (47%) are 17–27 residues long, whereas 93% of the low-resolution helices fall into this interval (gray line). Half of the low-resolution helices (50%) are 20–24 residues long, whereas 25% of the high-resolution helices fall into this interval (dashed line).

1, inset). Helices from 17–27 residues accounted for less than half of the high-resolution and for 93% of the low-resolution data. The distribution of lengths was clearly shifted toward longer helices in the high-resolution data (Fig. 1). In particular, 12 high-resolution helices (9%) were longer than 34 residues, that is, fall outside the range of what the low-resolution experiments suggested as possible lengths for membrane helices. The following four proteins had the longest helices: (1) the cytochrome BC complex (1BGY:D 34 residues, 1BGY:G 43 residues; Iwata et al. 1998); (2) the calcium ATPase (1EUL:A, TMH 6, 39 residues; Toyoshima et al. 2000); (3) the cytochrome C oxidase (2OCC:I, TMH 1, 39 residues; Tsukihara et al. 1996); and (4) the fumarate reductase (1FUM:C, TMH 2, 38 residues; Iverson et al. 1999). Typically, the long helices were either slightly bent (1BGY:D, 1fum:C) or extended into globular domains (1eul:A, Fig. 2). Overall, the recent high-resolution data appeared to strongly challenge the assumption of many developers of prediction methods, namely that the vast majority of membrane helices are 17–25 residues long. This incorrect assumption has been implemented as a more or less rigid constraint into most existing prediction methods. In fact, to implement such a constraint is important as many regions in membrane proteins consist of >40–60 consecutive hydrophobic residues that usually form more than one membrane helix. These long helices have to be “dissected” by prediction methods, not the least to accurately predict topology. Thus, the unexpected reality observed in high-resolution structures (Figs. 1, 2) complicates the prediction task.

**High-resolution structures revealed considerable proportion of short loops.** Monne, von Heijne, and coworkers experimentally established the propensities of amino acids to form tight turns (loops) between membrane helices (Monne et al. 1999a,b; Monne and von Heijne 2001). They find that the charged and polar amino acids DEQNRK as well as the flexible P and G have the highest preferences to form tight turns. However, in their data set these investigators found very few proteins with loops shorter than 7 residues. Plotting the length distribution of loops, we noticed two important results (Fig. 3): (1) Low-resolution experiments tended to suggest significantly longer loops than high-resolution structures, and (2) about half of all loops in high-resolution structures were 10 residues or shorter and >20% of the high-resolution loops were  $\leq 5$  residues long. Obviously, we cannot expect that the 36 sequence-unique high-resolution chains used in our study (see Materials and Methods) are fully representative for all helical membrane proteins. Given that we predict about 20,000 helical membrane proteins in the five entirely sequenced eukaryotes alone (Liu and Rost 2001, 2002), we also doubt that the 165-sequence-unique low-resolution proteins (see Materials and Methods) are more representative. Clearly, the high-resolution data are more accurate than the low-resolution data. Thus, our



**Fig. 2.** Long membrane helices in high-resolution structures. The plots were generated using the RASMOL program (Sayle and Milner-White 1995). All transmembrane helices shown in black extend over  $>38$  residues.

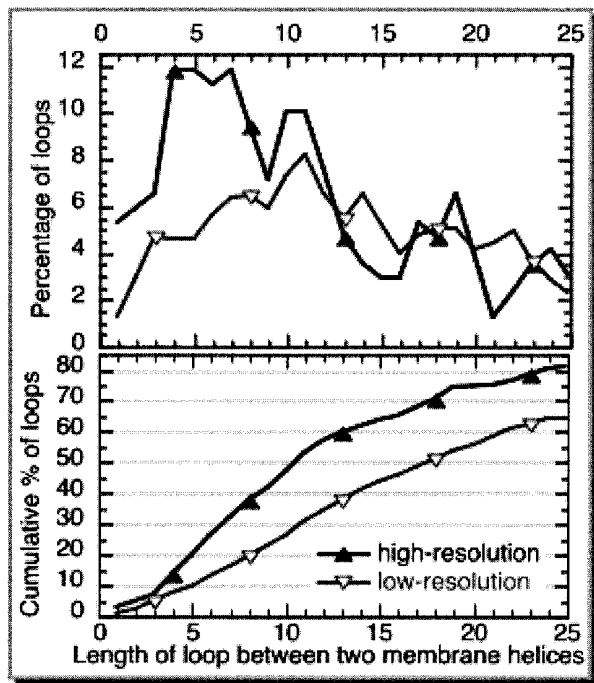
data suggested that a considerable percentage of all loops between membrane helices are very short.

#### *Long helices and short loops challenge prediction methods*

*Short loops predicted at lower accuracy.* As discussed previously, about half of all loops connecting two membrane helices are shorter than 10 residues. Most prediction methods compile averages using windows of 13–25 consecutive residues. Thus, the signal from the flanking helices may override that for a short loop. If so, we expect short loops to be predicted less accurately. Our data clearly confirmed this suspicion: Shorter loops are predicted by all methods less accurately than long loops (Fig. 4 and Table 1). The low-resolution data suggested that prediction accuracy decreased significantly for loops shorter than 10 residues, whereas the high-resolution data suggested the significant decrease to occur for loops shorter than 7 residues (Fig. 4). For example, although  $\sim 90\%$  of the loops longer than 15 residues were correctly detected by the advanced prediction methods,  $<60\%$  of the loops  $\leq 5$  residues were identified (Fig. 4,

left graph). These data suggested that methods that predict membrane helices have explicitly embedded loop preferences, such as the ones derived by the von Heijne group (Monne et al. 1999a, 1999b; Monne and von Heijne 2001).

*Prediction accuracy depended on helix length.* When we correlated prediction accuracy to the length of the observed TMH, we observed three overall trends (Fig. 5): (1) For any chosen threshold in the number of residues  $N$  with  $N \leq 32$  residues used to group membrane helices into short and long, helices shorter than  $N$  were predicted less accurately than were helices longer than  $N$ ; (2) the trend was inverted for helices longer than 32 residues (only available for high-resolution data). These very long helices were predicted less accurately than all other helices; and (3) helices shorter than 17–20 residues posed an even stronger challenge to prediction methods than shorter helices (a significant drop of accuracy is shown in Figure 5). At first sight, the decrease in prediction accuracy for helices longer than 32 residues may appear irrelevant in context of predicting membrane helical proteins for entire proteomes (Goffeau et al. 1993; Rost et al. 1996b; Arkin et al. 1997; Frishman and Mewes 1997; Jones 1998, Wallin and von Heijne 1998; Gupta et al. 1999;



**Fig. 3.** Length distributions for loops between two membrane helices. The lower graph gives the percentage of all loops between two membrane helices that have  $N$  (shown 0–25) residues; the upper graph shows the cumulative data for example, 65% of all loops in high-resolution structures (black lines with solid triangles) are  $\leq 15$  residues long, whereas 65% of the loops in low-resolution experiments are  $\leq 25$  residues. Significantly more short loops are observed in the high- than in the low-resolution data. Although  $\sim 40\%$  of the high-resolution loops were shorter than 9 residues, only half as many loops in the low-resolution set were that short.

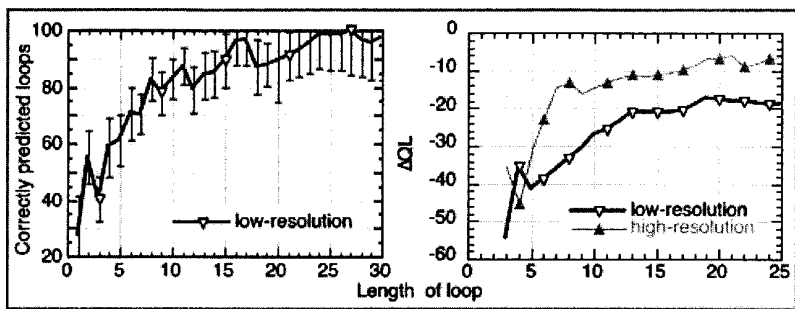
Krogh et al. 2001; Liu and Rost 2001). However, if we can generalize from the currently known high-resolution structures, we expect that  $\sim 20\%$  of all membrane helices are longer than 32 residues (Fig. 1). For the five entirely se-

quenced eukaryotic proteomes, this translates to  $\sim 5000$  proteins with a helix longer than 32 residues (Liu and Rost 2001, 2002).

#### *Detailed analysis of the mistakes in predicting long helices*

*Advanced methods miss long helices; simple methods incorrectly split them.* To explore why membrane protein prediction methods have trouble with long helices, we visually classified the predictions for long helices ( $\geq 33$  residues) as being (1) correct, (2) incorrectly cut into two membrane helices, and (3) not predicted at all (Fig. 6). Advanced methods are correct at predicting these long helices with an accuracy of  $>90\%$ . However, when these methods fail, it is about three times more likely to not predict a helix at all than it is to incorrectly predict the helix as two shorter helices. This may suggest that advanced methods mostly distinguish correctly between the membrane and the non-membrane parts of long helices. Simple hydrophobicity-based methods identified only  $\sim 71\%$  of the long helices correctly. In contrast to the advanced methods, the errors of simple methods had a six times higher rate of incorrectly splitting long helices than they had of missing the helix. This may suggest that the difficulty of simple methods with long helices is primarily due to overpredicting helical regions. This is supported by the fact that simple methods have great sensitivity but poor specificity at detecting TMH (Chen et al. 2002). In contrast, advanced methods have better specificity at detecting a TMH, as is indicated by their high accuracy of predicting even long helices. However, the price for being highly specific is that they can miss some TMH.

*Visual inspections of a few cases suggest to combine advanced and simple methods.* If one examines the cytochrome BC complex, one of the helices (residues 197–231 of 1BGY:D) was not predicted by 50% of the advanced



**Fig. 4.** Accuracy in predicting short loops. (Left) Percentage of loops with  $N$  (0–30) residues that were correctly predicted by all advanced prediction methods; the bars indicate the error estimates for these values. Note that the high-resolution data was too small to display noncumulative distributions. (Right) Difference in prediction accuracy between loops shorter and longer than the respective loop length (Eq. 1). All values were negative, implying that longer loops were always predicted at higher accuracy than shorter ones.

**Table 1.** Performance for short loops<sup>a</sup>

Method	High-resolution data		Low-resolution data	
	Q <sub>ok</sub> ±18	Q <sub>loop</sub> ±17	Q <sub>ok</sub> ±9	Q <sub>loop</sub> ±8
<i>ERROR</i>				
DAS	88	93	76	83
TopPred2	61	60	40	43
TMHMM1	61	55	24	38
PRED-TMR	50	49	45	53
SOSUI	49	68	26	38
PHDpsiHtm08	38	44	13	19
PHDhtm08	32	33	14	18
HMMTOP2	30	40	40	48
PHDhtm07	28	33	13	19
Wolfenden	91	90	54	72
Ben-Tal	48	52	46	63
KD	29	38	18	26
WW	28	44	26	34
GES	21	23	26	33
Sweet	21	22	21	30
A-Cid	20	28	9	16
Bull-Breese	20	21	19	23
Lawson	18	20	12	17
EM	12	11	16	25
Eisenberg	10	17	24	27
Nakashima	10	11	24	33
Heijne	10	11	16	24
Roseman	10	11	17	28
Levitt	10	11	13	18
Radzicka	10	10	11	20
Fauchere	9	10	14	23
Av-Cid	0	11	12	20
Hopp-Woods	0	0	15	23

<sup>a</sup> *Data set:* 36 sequence-unique high-resolution membrane helical proteins from PDB (Materials and Methods); 165 sequence-unique low-resolution membrane helical proteins from SWISS-PROT (Materials and Methods). *Methods:* advanced and hydrophobicity-based methods are separated by a blank row, abbreviations given in Materials and Methods. The advanced methods are sorted by alphabet, the simple hydrophobicity-based methods according to the Q<sub>ok</sub> score (Chen et al. 2002).

*Error:* The estimates for the standard error of the accuracy resulted from a bootstrap experiment with M = 100 and K = 18.

*Accuracy:* Loops were considered predicted correctly when at least one residue of the observed loop was predicted as loop. The actual value is the difference between the accuracy of all loops <7 residues and that for all loops ≥7 residues (Eq. 2).

*Numbers in italics:* Two standard errors below the numerically highest value in each column.

*Note of caution:* All methods are tested on the same set of proteins. However, the numbers are NOT from a cross-validation experiment, i.e., some methods may have used some of the proteins for training. Generally, newer methods are more likely to be overestimated than older ones.

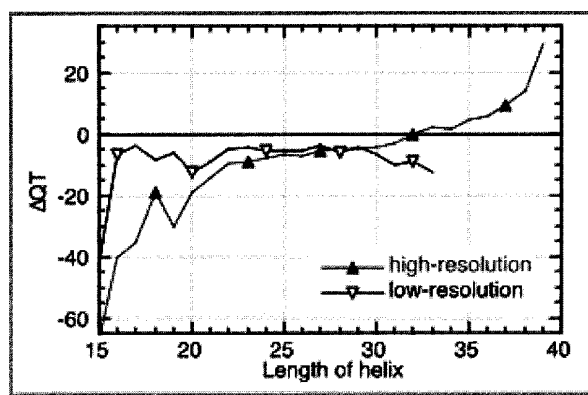
methods. One explanation for the difficulty in detecting this particular helix is the lack of a consecutive stretch of at least 17 hydrophobic residues. Even when the methods did predict its presence, they failed to predict residues at the amino and carboxyl termini of the helix (6 residues or more at either end). The residues not predicted as TM were often either polar or charged amino acids. For instance, residues 197–203 are EHDHRKR and residues 223–231 are

KRHKWSVLK. This theme of predicting the core hydrophobic region of a TMH but not detecting the more polar amino- and carboxyl-ends for long helices was repeated for most of the predictions by the advanced methods. The simple hydrophobic methods tended to identify these long helices. For instance, for 1BGY:D, many of the simple hydrophobicity-based methods missed the first 7 but correctly detected the last 7 residues. In fact three simple methods captured the entire length of the observed helix. This example might suggest a potential strategy to deal with long membrane helices: (1) Identify consensus regions for long membrane helices through simple hydrophobicity scales; (2) determine the core of the membrane segment through advanced prediction methods; and (3) extend the predicted helix in the directions of both the amino and carboxyl termini by using a scoring matrix that is optimized for non-TM residues and by setting the boundaries of extension within the region defined by the simple methods.

## Materials and methods

*Data sets.* For the high-resolution data, we started with 105 chains from helical membrane proteins with high-resolution structures deposited in PDB (Berman et al. 2000). We then reduced the bias in this dataset resulting from multiple copies of similar proteins. This left a set of 36 high-resolution proteins that were sequence-unique in the sense that no pair in that list had an HSSP distance above 0. (Rost 1999; for more details, see Chen et al. 2002.) We identified membrane regions through DSSP (Kabsch and Sander 1983). For the low-resolution data, we used a sequence-unique subset of the expert-curated set of helical membrane proteins for which good low-resolution experimental evidence about localization was available (Moller et al. 2000). The final sequence-unique subset contained 165 proteins.

*Advanced prediction methods.* We referred to prediction methods as advanced when they implement more than simple hydrophobicity scales. We tested the following programs: DAS,



**Fig. 5.** Accuracy in predicting long membrane helices. The difference in prediction accuracy between membrane helices shorter and longer than N residues (Eq. 2) is shown. Negative values imply that short helices were predicted less accurately than longer ones.

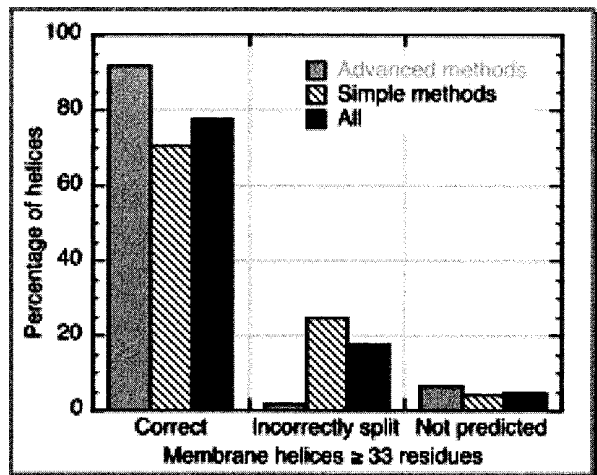


Fig. 6. Visual inspection of errors for long membrane helices. Prediction methods incorrectly split long helices ~17% and miss them 5% of the time. When advanced methods predict long helices incorrectly, it is about three times more likely to not predict a helix at all than to split it. In contrast, simple hydrophobic methods are six times more likely to incorrectly predict a long helix as two shorter ones than to not predict the helix at all.

HMMTOP (version 2), PHDhtm, PHDpsihm, PRED-TMR, SOSUI, TMHMM (version 2), and TopPred2. TopPred2 averages the GES-scale of hydrophobicity (Engelman et al. 1986) using a trapezoid window (von Heijne 1992; Sipos and von Heijne 1993). PHDhtm combines a neural network using evolutionary information with a dynamic programming optimization of the final prediction (Rost et al. 1995, 1996b). PHDpsihm uses PSI-BLAST (Altschul et al. 1997) alignments as input (B. Rost, unpubl.). DAS optimizes the use of hydrophobicity plots (Cserző et al. 1997). SOSUI (Hirokawa et al. 1998) uses a combination of hydrophobicity and amphiphilicity preferences to predict membrane helices. TMHMM is the most advanced—and seemingly most accurate—current method to predict membrane helices (Sonnhammer et al. 1998). It embeds a number of statistical preferences and rules into a hidden Markov model to optimize the prediction of the localization of membrane helices and their orientation. (Note: Similar concepts are used for HMMTOP; Tusnady and Simon 1998). PRED-TMR uses a standard hydrophobicity analysis with emphasis on detecting the ends and beginnings of membrane helices (Pasquier et al. 1999).

*Simple methods exclusively based on hydrophobicity scales.* We also implemented our in-house prediction methods that simply used various hydrophobicity scales for prediction. In particular, we tested the following scales: A-Cid, normalized hydrophobicity scale for  $\alpha$  proteins (Cid 1992); Av-Cid, normalized average hydrophobicity scale (Cid 1992); Ben-Tal, hydrophobicity scale representing free energy of transfer of an amino acid from water into the center of the hydrocarbon region of a model lipid bilayer (Kessel and Ben-Tal 2002); Bull-Breese, Bull-Breese hydrophobicity scale (Bull 1974); Eisenberg, normalized consensus hydrophobicity scale (Eisenberg et al. 1984); EM, solvation-free energy (Eisenberg and McLachlan 1986); Fauchere, hydrophobic parameter  $\pi$  from the partitioning of *N*-acetyl-amino acid amides (Fauchere and Pliska 1983); GES, hydrophobicity property (Engelman et al. 1986); Heijne, transfer-free energy to lipophilic phase (von Heijne and Blomberg 1979); Hopp-Woods, Hopp-Woods hydro-

philicity value (Hopp and Woods 1981); KD, Kyte-Doolittle hydrophobicity index (Kyte and Doolittle 1982); Lawson, transfer-free energy (Lawson et al. 1984); Levitt, hydrophobic parameter (Levitt 1976); Nakashima, normalized composition of membrane proteins (Nakashima et al. 1990); Radzicka, transfer-free energy from 1-octanol to water (Radzicka and Wolfenden 1988); Roseman, solvation-corrected side chain hydrophobicity (Roseman 1988); Sweet, optimal matching hydrophobicity (Sweet and Eisenberg 1983); Wolfenden, hydration potential (Wolfenden et al. 1981); and WW, Wimley-White scale (Jayasinghe et al. 2001). Replacing the WW scale with each of the above-mentioned hydrophobicity indices, we used the WW algorithm to evaluate the predictive performance of each index.

*Measuring accuracy.* To establish whether or not short loops and long membrane helices pose particular problems for prediction methods, we have to deviate from the scores used to evaluate performance of membrane prediction methods (Chen et al. 2002). In particular, we introduced the following scores that describe the difference in performance between short and long loops ( $\Delta Q_L(N)$ , Eq. 1), and that between short and long TMH ( $\Delta Q_T(N)$ , Eq. 2).

(1) *Short loops.* We evaluated the performance of predicting short loops, that is, regions connecting two membrane helices with  $\leq N$  residues by compiling the difference between the accuracy in predicting short and long loops:

$$\Delta Q_L(N) = 100 \cdot \left\{ \frac{N_{loop < N \text{ identified}}}{N_{loop < N \text{ observed}}} - \frac{N_{loop \geq N \text{ identified}}}{N_{loop \geq N \text{ observed}}} \right\} \quad (\text{Eq. 1})$$

where  $N$  is the number of residues;  $N_{loop < N \text{ identified}}$  is the number of loops with  $< N$  residues that were correctly predicted, and  $N_{loop < N \text{ observed}}$ , the number of loops observed to have  $< N$  residues. We considered a loop of  $N$  residues to be correctly predicted if at least 1 residue in that loop was predicted, that is, if the presence of a break between two helices was correctly identified.  $\Delta Q_L(n)$  could adopt values between  $-100$  and  $100$ ; negative values indicate that longer loops are predicted more accurately than shorter ones.

(2) *Long helices.* In analogy to the score describing the performance for short loops, we evaluated the performance of predicting long TMH by compiling the difference between the accuracy in predicting short and long helices:

$$\Delta Q_T(N) = 100 \cdot \left\{ \frac{N_{tm < N \text{ identified}}}{N_{tm < N \text{ observed}}} - \frac{N_{tm \geq N \text{ identified}}}{N_{tm \geq N \text{ observed}}} \right\} \quad (\text{Eq. 2})$$

where  $N_{tm \geq N \text{ identified}}$  is the number of TMH with  $\geq N$  residues that were correctly predicted and  $N_{tm \geq N}$ , the number of TMH with  $\geq N$  residues observed. We considered a helix to be correctly predicted if it overlapped at least for 3 residues with the observed helix and if it was predicted as one continuous helix (over the region of the observed helix). This measure is illustrated in the following example for a prediction ( $T = \text{TM}$ ;  $\neq$  loop):

Observed: -----TTTTTTTTTTTTTTTTTTTT-----  
 Predict 1: -----TTTTTTTTTT-----  
 Predict 2: ---TTTTTTTTTTTTTT-TTTTTTTTTTTTTTT-----

In this example, Predict 1 is right and Predict 2 is wrong because all we are trying to capture is whether or not methods tended to split long TMH.  $\Delta Q_T(N)$  ranges from  $-100$  to  $100$ ; it becomes negative if helices shorter than  $N$  residues are predicted more accurately than helices  $\geq N$ .

## Acknowledgments

Thanks to Jinfeng Liu (Columbia) for computer assistance and the collection of genome datasets. The work of B.R. was supported by grants 1-P50-GM62413-01 and RO1-GM63029-01 from the National Institutes of Health (NIH) and by grant DBI-0131168 from the National Science Foundation (NSF). Last, but not least, thanks to all those who deposit their experimental data in public databases and to those who maintain these databases.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

## References

- Altschul, S., Madden, T., Shaffer, A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. 1997. Gapped Blast and PSI-Blast: A new generation of protein database search programs. *Nucleic Acids Res.* **25**: 3389–3402.
- Arkin, I.T., Brünger, A.T., and Engelman, D.M. 1997. Are there dominant membrane protein families with a given number of helices? *Proteins* **28**: 465–466.
- Bairoch, A. and Apweiler, R. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28**: 45–48.
- Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. 2000. The Protein Data Bank. *Nucleic Acids Res.* **28**: 235–242.
- 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.
- Bowie, J.U. 1997. Helix packing in membrane proteins. *J. Mol. Biol.* **272**: 780–799.
- Bull, H.B. and Breese, K. 1974. Surface tension of amino acid solutions: A hydrophobicity scale of the amino acid residues. *Arch. Biochem. Biophys.* **161**: 665–670.
- Casadio, R., Farielli, P., Taroni, C., and Compiani, M. 1996. A predictor of transmembrane  $\alpha$ -helix domains of proteins based on neural networks. *Eur. J. Biophys.* **24**: 165–178.
- Chen, C.P., Kerytsky, A., and Rost, B. 2002. Transmembrane helix predictions revisited. *Protein Sci.* (this issue).
- Cid, H., Bunster, M., Canales, M., and Gazitua, F. 1992. Hydrophobicity and structural classes in proteins. *Prot. Engin.* **5**: 373–375.
- Cserző, M., Wallin, E., Simon, I., von Heijne, G., and Elofsson, A. 1997. Prediction of transmembrane  $\alpha$ -helices in prokaryotic membrane proteins: The dense alignment surface method. *Prot. Engin.* **10**: 673–676.
- Eisenberg, D. and McLachlan, A.D. 1986. Solvation energy in protein folding and binding. *Nature* **319**: 199–203.
- Eisenberg, D., Weiss, R.M., and Terwilliger, T.C. 1984. The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc. Natl. Acad. Sci.* **81**: 140–144.
- Engelman, D.M., Steitz, T.A., and Goldman, A. 1986. Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annu. Rev. Biophys. Chem.* **15**: 321–353.
- Fauchere, J.L. and Pliska, V. 1983. Hydrophobic parameters  $\pi$  of amino-acid side chains from the partitioning of N-acetyl-amino-acid amides. *Eur. J. Med. Chem.* **18**: 369–375.
- Frishman, D. and Mewes, H.W. 1997. Protein structural classes in five complete genomes. *Nature Struct. Biol.* **4**: 626–628.
- Goffeau, A., Nakai, K., Slonimski, P., and Risler, J.-L. 1993. The membrane proteins encoded by yeast chromosome III genes. *FEBS Lett.* **325**: 112–117.
- Gupta, R., Jung, E., Gooley, A.A., Williams, K.L., Brunak, S., and Hansen, J. 1999. Scanning the available Dictyostelium discoideum proteome for O-linked GlcNAc glycosylation sites using neural networks. *Glycobiology* **9**: 1009–1022.
- Hirokawa, T., Boon-Chieng, S., and Mitaku, S. 1998. SOSUI: Classification and secondary structure prediction system for membrane proteins. *Bioinformatics* **14**: 378–379.
- Hopp, T.P. and Woods, K.R. 1981. Prediction of protein antigenic determinants from amino acid sequences. *Proc. Natl. Acad. Sci.* **78**: 3824–3828.
- Ikeda, M., Arai, M., Lao, D.M., and Shimizu, T. 2001. Transmembrane topology prediction methods: A reassessment and improvement by a consensus method using a dataset of experimentally-characterized transmembrane topologies. *Silico Biol.* **1**: <http://www.bioinfo.de/isb/2001/2002/0003/>.
- Iverson, T.M., Luna-Chavez, C., Cecchini, G., and Rees, D.C. 1999. Structure of the *E. coli* fumarate reductase respiratory complex. *Science* **284**: 1961.
- Iwata, S., Lee, J.W., Okada, K., Lee, J.K., Iwata, M., Rasmussen, B., Link, T.A., Ramaswamy, S., and Jap, B.K. 1998. Complete structure of the 11-subunit bovine mitochondrial cytochrome bc<sub>1</sub> complex. *Science* **281**: 64–71.
- Jayasinghe, S., Hristova, K., and White, S.H. 2001. Energetics, stability, and prediction of transmembrane helices. *J. Mol. Biol.* **312**: 927–934.
- Jones, D.T. 1998. Do transmembrane protein superfolds exist? *FEBS Lett.* **423**: 281–285.
- 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. and Sander, C. 1983. Dictionary of protein secondary structure: Pattern recognition of hydrogen bonded and geometrical features. *Biopolymers* **22**: 2577–2637.
- Kessel, A. and Ben-Tal, N. 2002. Free energy determinants of peptide association with lipid bilayers. In *Peptide-lipid interactions* (eds. S. Simon and T. McIntosh). Academic Press, San Diego, CA (in press).
- Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E.L. 2001. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *J. Mol. Biol.* **305**: 567–580.
- Kyte, J. and Doolittle, R.F. 1982. A simple method for displaying the hydrophobic character of a protein. *J. Mol. Biol.* **157**: 105–132.
- Lawson, E.Q., Sadler, A.J., Harmatz, D., Brandau, D.T., Micanovic, R., MacElroy, R.D., and Middaught, C.R. 1984. A simple experimental model for hydrophobic interactions in proteins. *J. Biol. Chem.* **259**: 2910–2912.
- Levitt, M. 1976. A simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol.* **104**: 59–107.
- Liu, J. and Rost, B. 2001. Comparing function and structure between entire proteomes. *Protein Sci.* **10**: 1970–1979.
- . 2002. Target space for structural genomics revisited. *Bioinformatics* **18**: 922–933.
- Möller, S., Kriventseva, E.V., and Apweiler, R. 2000. A collection of well characterised integral membrane proteins. *Bioinformatics* **16**: 1159–1160.
- Möller, S., Croning, D.R., and Apweiler, R. 2001. Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics* **17**: 646–653.
- Monne, M. and von Heijne, G. 2001. Effects of 'hydrophobic mismatch' on the location of transmembrane helices in the ER membrane. *FEBS Lett.* **496**: 96–100.
- Monne, M., Hermansson, M., and von Heijne, G. 1999a. A turn propensity scale for transmembrane helices. *J. Mol. Biol.* **288**: 141–145.
- Monne, M., Nilsson, I., Elofsson, A., and von Heijne, G. 1999b. Turns in transmembrane helices: Determination of the minimal length of a "helical hairpin" and derivation of a fine-grained turn propensity scale. *J. Mol. Biol.* **293**: 807–814.
- Nakashima, H., Nishikawa, K., and Ooi, T. 1990. Distinct character in hydrophobicity of amino acid composition of mitochondrial proteins. *Proteins* **8**: 173–178.
- Pasquier, C., Promponas, V.J., Palaios, G.A., Hamodrakas, J.S., and Hamodrakas, S.J. 1999. A novel method for predicting transmembrane segments in proteins based on a statistical analysis of the SwissProt database: the PRED-TMR algorithm. *Protein Eng.* **12**: 381–385.
- Persson, B. and Argos, P. 1996. Topology prediction of membrane proteins. *Protein Sci.* **5**: 363–371.
- Radzicka, A. and Wolfenden, R. 1988. Comparing the polarities of the amino acids: Side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution. *Biochem.* **27**: 1664–1670.
- Roseman, M.A. 1988. Hydrophilicity of polar amino acid side-chains is markedly reduced by flanking peptide bonds. *J. Mol. Biol.* **200**: 513–522.
- Rost, B. 1996. PHD: Predicting one-dimensional protein structure by profile based neural networks. *Methods Enzymol.* **266**: 525–539.
- . 1999. Twilight zone of protein sequence alignments. *Protein Eng.* **12**: 85–94.
- . 2001. Protein secondary structure prediction continues to rise. *J. Struct. Biol.* **134**: 204–218.

- Rost, B., Casadio, R., Fariselli, P., and Sander, C. 1995. Prediction of helical transmembrane segments at 95% accuracy. *Protein Sci.* **4**: 521–533.
- Rost, B., Casadio, R., and Fariselli, P. 1996a. Refining neural network predictions for helical transmembrane proteins by dynamic programming. In *Fourth International Conference on Intelligent Systems for Molecular Biology* (eds. D. States), pp. 192–200. AAAI Press, St. Louis, MO, Menlo Park, CA.
- . 1996b. Topology prediction for helical transmembrane proteins at 86% accuracy. *Protein Sci.* **5**: 1704–1718.
- Sayle, R.A. and Milner-White, E.J. 1995. RASMOL: Biomolecular graphics for all. *Trends Biochem. Sci.* **20**: 37.
- Sipos, L. and von Heijne, G. 1993. Predicting the topology of eukaryotic membrane proteins. *Eur. J. Biochem.* **213**: 1333–1340.
- Sonnhammer, E.L.L., von Heijne, G., and Krogh, A., 1998. A hidden Markov model for predicting transmembrane helices in protein sequences. In *Sixth International Conference on Intelligent Systems for Molecular Biology (ISMB98)* (eds. J. Glasgow), pp. 175–182. AAAI Press, Montreal, Canada.
- Sweet, R.M. and Eisenberg, D. 1983. Correlation of sequence hydrophobicities measures similarity in three-dimensional protein structure. *J. Mol. Biol.* **171**: 479–488.
- Toyoshima, C., Nakasako, M., Nomura, H., and Ogawa, H. 2000. Crystal structure of the calcium pump of sarcoplasmic reticulum at 2.6 Å resolution. *Nature* **405**: 647.
- Tsukihara, T., Aoyama, H., Yamashita, E., Tomizaki, T., Yamaguchi, H., Shinzawa-Itoh, K., Nakashima, R., Yaono, R., and Yoshikawa, S. 1996. The whole structure of the 13-subunit oxidized cytochrome C oxidase at 2.8 Å. *Science* **272**: 1136.
- Tusnady, G.E. and Simon, I. 1998. Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *J. Mol. Biol.* **283**: 489–506.
- . 2001. Topology of membrane proteins. *J. Chem. Inf. Comput. Sci.* **41**: 364–368.
- von Heijne, G. 1992. Membrane protein structure prediction. *J. Mol. Biol.* **225**: 487–494.
- . 1996. Prediction of transmembrane protein topology. In *Protein structure prediction* (ed. M.J. E. Sternberg), pp. 101–110. Oxford Univ. Press., Oxford, UK.
- von Heijne, G. and Blomberg, C. 1979. Trans-membrane translocation of proteins: The direct transfer model. *Eur. J. Biochem.* **97**: 175–181.
- Wallin, E. and von Heijne, G. 1998. Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci.* **7**: 1029–1038.
- Wolfenden, R., Andersson, L., Cullis, P.M., and Southgate, C.C.B. 1981. Affinities of amino acid side chains for solvent water. *Biochemistry* **20**: 849–855.