Eur. J. Biochem. *213*, 1333–1340 (1993)
© FEBS 1993

Predicting the topology of eukaryotic membrane proteins

Laszlo SIPOS[1] and Gunnar VON HEIJNE[2]

[1] Research Group for Theoretical Biophysics, Department of Theoretical Physics, Royal Institute of Technology, Stockholm, Sweden
[2] Department of Molecular Biology, Karolinska Institute Center for Structural Biochemistry, NOVUM, Huddinge, Sweden

(Received January 15/February 22, 1993) – EJB 93 0071/2

We show that the so-called 'positive inside' rule, i.e. the observation that positively charged amino acids tend to be more prevalent in cytoplasmic than in extra-cytoplasmic segments in transmembrane proteins [von Heijne, G. (1986) *EMBO J. 5*, 3021–3027], seems to hold for all polar segments in multi-spanning eukaryotic membrane proteins irrespective of their position in the sequence and hence can be used in conjunction with hydrophobicity analysis to predict their transmembrane topology. Further, as suggested by others, we confirm that the net charge difference across the first transmembrane segment correlates well with its orientation [Hartmann, E., Rapoport, T. A. and Lodish, H. F. (1989) *Proc. Natl Acad. Sci. USA 86*, 5786–5790], and that the overall amino-acid composition of long polar segments can also be used to predict their cytoplasmic or extra-cytoplasmic location [Nakashima, H. and Nishikawa, K. (1992) *FEBS Lett. 303*, 141–146]. We present an approach to the topology prediction problem for eukaryotic membrane proteins based on a combination of these methods.

Integral membrane proteins seem to come in two basic varieties: the helical bundle proteins with membrane domains built from a bundle of hydrophobic transmembrane α helices [1, 2], and the β-barrel proteins which form large, anti-parallel β barrels with a hydrophobic outer surface facing the lipids [3, 4]. In both cases, the requirement that a continuous outer apolar surface must be formed has necessitated a structurally rather well-defined solution. This greatly facilitates the task of predicting structure from sequence, and the secondary structure of both types of protein can now be fairly well predicted [4, 5].

For the helical bundle proteins, two easily identified features of the nascent chain seem to be the major structural determinants: the long, apolar stretches that form the transmembrane α helices and the biased distribution of arginines and lysines in the polar regions exposed on either side of the membrane, the 'positive inside' rule [6, 7]. In fact, we have recently shown that transmembrane segments in bacterial inner-membrane proteins can be predicted with very high confidence if this charge bias is taken into account [5].

The possibility of similarly improving the prediction of transmembrane helices in eukaryotic membrane proteins has not been explored so far. *A priori*, there are a number of reason why a charge-bias analysis of the kind shown to work well for prokaryotic proteins may be less informative in this case: in contrast to the situation in prokaryotes, the initial insertion of eukaryotic proteins into the membrane of the endoplasmic reticulum is thought to be largely co-translational, suggesting that only the most N-terminal apolar segment(s) will determine the final topology; the (Arg+Lys) bias between cytoplasmic and extra-cytoplasmic (lumenal) segments is less extreme than in prokaryotes [7]; it appears that the net charge difference (including negatively charged Asp and Glu residues) across the first transmembrane segment, rather than the difference in the number of positively charged residues alone, correlates best with the final topology [8]. Experimental information on the topological role of charged residues in eukaryotic proteins is limited to cases with a single N-terminal transmembrane segment, and it is clear that both positively and negatively charged residues on both sides of the hydrophobic stretch can influence the final orientation of such proteins [9–11], in conformity with the 'positive inside' [6] and 'charge difference' [8] rules.

Nevertheless, we now show that the 'positive inside' rule seems to hold well for all polar segments of eukaryotic plasma membrane proteins shorter than approximately 60 residues (although the bias appears to become progressively less marked for more C-terminal segments), and can thus be used for prediction purposes. Further, we also show that methods based on weak but significant compositional differences between long cytoplasmic and extra-cytoplasmic segments [12] can be profitably built into the prediction scheme and can give topological information on segments that are too long to be useful in the charge-bias analysis. Finally, we discuss some limitations of current prediction methods and identify parameters that have not yet been determined with sufficient accuracy.

METHODS

Protein database

16 eukaryotic multi-spanning membrane proteins with partially or fully characterized topologies were analyzed (Table 1). The topologies were derived from the following

Correspondence to G. von Heijne, Department of Molecular Biology, Karolinska Institute Center for Structural Biochemistry, NOVUM, S-141-57 Huddinge, Sweden

Fax: +46 8 774 55 38.

Abbreviation. TRAM, translocating chain-associating membrane protein.

experimental data and hydrophobicity plots, as described previously [7].

SEC63

Fusion-protein analysis suggests that the C-terminus and a region between the first and second putative transmembrane segments are cytoplasmic and that a region between the second and third putative transmembrane segments is lumenal. Protease accessibility confirms the cytoplasmic location of the C-terminus [13].

SEC62

Fusion protein analysis suggests that the N-terminus and C-terminus are cytoplasmic and that a region between the first and second putative transmembrane segments is lumenal. Protease accessibility confirms the cytoplasmic location of the N-terminus and C-terminus [14]. The experimental data do not, however, rule out the possibility that a mildly hydrophobic segment very near the N-terminus (residues 16−36) also spans the membrane with the N-terminus in the lumen, as predicted by the 'charge difference' rule. Our general conclusions are not affected if this is indeed the case.

Adenovirus E3 10.4 K protein

Protease accessibility suggests that the C-terminus is cytoplasmic [15].

Gastric H^+/K^+ ATPase, α-subunit

The N-terminus and a large region between residues 325 and 800 are cytoplasmic as assayed by protease accessibility [16−18]. The location of the N-terminus is also supported by antibody binding [19]. There are four well-defined hydrophobic stretches in the N-terminal 325 residues, and this part is included in our statistics. The topology of the C-terminal part of the molecule is still controversial, and this domain was not included in the present study.

Mouse-hepatitis-virus E1 protein

The N-terminus is glycosylated and hence lumenal, while the C-terminus is cytoplasmic as assayed by protease accessibility in microsomes and by the observation that it interacts with the nucleocapsid [20].

Avian-infectious-bronchitis-virus E1 protein

The N-terminus is glycosylated and hence lumenal, while the C-terminus is cytoplasmic [21].

CD20 receptor

The C-terminus is cytoplasmic as indicated by antibody mapping [22].

Gap-junction β_1 protein

Antibody mapping and phosphorylation of the C-terminal domain suggest that the N-terminus, the C-terminus and a region between putative transmembrane segments 2 and 3 are cytoplasmic [23, 24]. Homology with the α_1 subunit suggests the same topology [25].

Multi-drug resistance protein

Antibody mapping suggests that the N-terminus and C-terminus and the large ATP-binding domains are cytoplasmic [26−28]. A region between the first and second putative transmembrane segments is glycosylated and hence lumenal [29]. We have only included the N-terminal half of the molecule (putative transmembrane segments 1−6) in our sample, since the first and second halves are homologous to each other.

Vacuolar H^+ ATPase proteolipid homologue

Homology to *Escherichia coli* ATPase F_0 subunit c suggests a topology with lumenal N-termini and C-termini and two cytoplasmic loops [30].

Acetylcholine receptor, α-subunit

The presence of an N-terminal, cleavable signal sequence (residues 1−24) places the N-terminus of the mature protein in the lumen. Known glycosylation sites, antibody mapping and fusion-protein analysis suggest four transmembrane domains, with the C-terminus located in the lumen [31−33].

STE2

Phosphorylation indicates that the C-terminal domain is cytoplasmic [34]. Fusion protein analysis suggests the presence of seven transmembrane segments, with the N-terminus lumenal [35].

3-Hydroxy-3-methylglutaryl coenzyme A reductase

Antibody mapping [36], analysis of glycosylation sites [37] and fusion-protein analysis [38] suggest that residues 1−345 form eight transmembrane segments, with both the N-terminus and C-terminus in the cytoplasm.

Rhodopsin

A large body of evidence supports the seven-transmembrane-segment model with the N-terminus lumenal and the C-terminus cytoplasmic [39].

Translocating chain-associating membrane protein (TRAM)

Proteolysis suggests that the C-terminus is cytoplasmic, and glycosylation suggests that the loop between the first and second putative transmembrane segments is lumenal [40].

Plasma-membrane H^+ ATPase

Protease digestion and sequencing of tryptic peptides have established the cytoplasmic location of five polar segments [41]. Together with the hydrophobicity profile, this suggests a topology with 10 transmembrane segments and both the N-terminus and C-terminus in the cytoplasm.

Prediction methods

Hydrophobicity and charge-bias analysis was performed as described in [5], i.e., using the GES hydrophobicity scale [42] and a trapezoid sliding window where the 11 central residues are all given equal weights, whereas the five flanking residues on either side are given progressively lower weights; this serves to reduce the noise in the hydrophobicity profile and is physically more realistic than a simple rectangular window, see [5]. Since charge pairing between oppositely charged residues located at positions $i,i+3$ or $i,i+4$ in a transmembrane helix may lower the free energy of membrane insertion relative to unpaired charges [43], such charge pairs were made 20.9 kJ/mol (5 kcal/mol) less unfavorable (this value is a rough guess, but does not seem to be unreasonable). From the hydrophobicity profile, candidate transmembrane segments were extracted automatically by a procedure that first identifies the highest peak, records its average hydrophobicity $\langle H \rangle$, and removes the part of the profile that corresponds to the 21 residues in this segment plus two additional flanking residues. This is repeated until either no stretch longer than 22 residues remains, or until no peak higher than the lower cutoff ($\langle H \rangle = 0.6$) remains. All peaks with $\langle H \rangle \geq 1.0$ were considered certain transmembrane segments, and all those with $0.6 \leq \langle H \rangle < 1.0$ were considered putative candidates. Finally, all possible topologies that included the certain transmembrane segments and either included or excluded each of the putative candidate segments were automatically generated, and the difference in the number of positively charged amino acids (Arg+Lys) between the two sides of each structure was calculated (see [5] for details). Also, an additional positive charge representing the free N-terminal amino group was added to the first polar segment. Polar segments longer than 60 residues were not included in the charge-bias calculation, since their content of charged residues does not seem to be dependent on their cytoplasmic or extra-cytoplasmic location. The possible topologies were then ranked in decreasing order of charge bias, and their orientation was predicted as the one with the more highly charged side facing the cytoplasm.

The net charge difference, $\Delta(\text{N}-\text{C}) = (n_\text{R}+n_\text{K}-n_\text{D}-n_\text{E})_\text{N-term} - (n_\text{R}+n_\text{K}-n_\text{D}-n_\text{E})_\text{C-term}$, between the 15 N-terminal and 15 C-terminal residues flanking the most N-terminal transmembrane segment was also calculated, as the sign of this quantity is know to correlate with the orientation of the transmembrane segment [8]: $\Delta(\text{N}-\text{C}) > 0$ implies a cytoplasmic location for the N-terminus; $\Delta(\text{N}-\text{C}) < 0$ an extra-cytoplasmic location.

Long segments (more than 60 residues) were analyzed by the compositional distance method [12]. Briefly, amino-acid frequencies were calculated for each segment, and the distances to the average amino-acid frequencies determined for reference samples of cytoplasmic and extra-cytoplasmic domains were determined. Finally, the difference between these two distances, $d_{\text{cyt}-\text{ext}}$, was calculated; a negative value indicates a preference for a cytoplasmic location. Although the authors of the original paper recommend that the reference values are taken from their Table 1, we found that those of their Table IV gave better results on the proteins listed in Table 1 and have used these throughout.

A program called TOP-PRED 2.0 (TOPology PREDiction program, written in THINK Pascal for Macintosh computers) implementing these methods is available upon request.

Amino-acid frequencies in various segments were compared using χ^2 analysis.

RESULTS AND DISCUSSION

The 'positive inside' rule holds for all short exposed segments

The basic strategy that was implemented in our previously described prediction scheme for prokaryotic membrane proteins [5] proceeds as follows. A standard hydrophobicity analysis is performed that identifies certain and putative transmembrane segments with average hydrophobicities above an upper cutoff H_1 or between H_1 and a lower cutoff H_2, a list of all possible candidate structures built from the certain segments and including all possible combinations of the putative segments is generated, and this list is ranked with the structure having the highest bias in the number of Arg+Lys taken as the best prediction (segments longer than 70 residues were not counted in this analysis; we have recently found that 60 residues may be a better cutoff for bacterial proteins [44]).

Obviously, in order for this to work, an appreciable bias in the distribution of Arg+Lys between cytoplasmic and extra-cytoplasmic polar segments must be present throughout the protein, not just around the most N-terminal transmembrane segment. We thus analyzed the Arg+Lys bias in cytoplasmic and extra-cytoplasmic segments progressively further removed from the N-terminus in a set of 16 eukaryotic plasma-membrane proteins with experimentally well-characterized topologies, (Table 1). The results are shown in Fig. 1A; a statistically significant ($P < 0.05$) bias is observed for the first three segments, with the cytoplasmic segments having on average a 2–3-fold higher frequency of Arg+Lys (fourfold in prokaryotes [6]). The bias, although always of the same sign, seems to get less pronounced as one moves away from the N-terminus, but the small number of proteins with more than five transmembrane segments in our sample makes it impossible to decide whether this progressive weakening is statistically significant. Further, there is no bias observed for segments longer than some 60 residues (Fig. 1B). It would thus appear that the same prediction strategy can be applied to both prokaryotic and eukaryotic membrane proteins, though the improvement from charge-bias analysis is expected to be somewhat less for the eukaryotic ones.

The distribution of Asp and Glu across the N-terminal transmembrane segment is biased (the 'charge difference' rule)

It has been suggested that the net charge difference ($n_\text{Arg}+n_\text{Lys}-n_\text{Asp}-n_\text{Glu}$) between the 15 residues flanking the most N-terminal transmembrane segment determines its orientation [8]. Indeed, the distribution of negatively charged residues in the first cytoplasmic and extra-cytoplasmic segments is consistent with this (Fig. 2): the extra-cytoplasmic segment is about twofold enriched in Asp+Glu ($P < 0.05$). This is only true for the most N-terminal segment, however, and there are no significant differences for the following segments. The 'charge difference' rule may thus only help to predict the orientation of the N-terminal segment and cannot be used to choose between different models for the more C-terminal parts.

Table 1. Proteins included in this study. n_{K+R} is the number of Lys+Arg residues in the segment (the N-terminal amino group is also counted). High n_{K+R} predicts a cytoplasmic location. $D_{cyt-ext}$ is the difference between the compositional distance from the segment to the average cytoplasmic (cyt) and extra-cytoplasmic (ext) amino-acid compositions ($D_{cyt-ext} < 0$ predicts a cytoplasmic location; see Methods). Δ(N−C) is the net charge difference between the 15 N-terminal and 15 C-terminal residues flanking the first transmembrane stretch [Δ(N− C) >0 predicts a cytoplasmic location of the N-terminus; see Methods]. The GenBank or Swiss-Prot accession number is also given. The experimental data do not rule out the possibility that a mildly hydrophobic, N-terminal segment of SEC62 (residues 16−36) also spans the membrane with the N-terminus facing the lumen. In this case, Δ(N−C) = −6 (i.e. it would correctly predict a lumenal N-terminus). ADV, adenovirus; MHV, mouse hepatitis virus; IBV, avian infectious bronchitis virus; MDR, multi-drug resistance; AcChR, acetylcholine receptor; HMG, 3-hydroxy-3-methylglutaryl.

Protein	Segment	Location	n_{K+R}	$D_{cyt-ext}$	Δ(N−C)	Accesion no.
SEC63	1− 17	ext	1	−	−2	X16388
	29− 96	cyt	13	−	−	
	108−223	ext	−	−0.46	−	
	235−663	cyt	−	−0.14	−	
SEC62	1−163	cyt	−	−1.31	−5	X51666
	175−190	ext	3	−	−	
	202−283	cyt	−	−0.40	−	
ADV E3 10.4K	1− 10	cyt	2	−	+3	J01917
	20− 45	ext	0	−	−	
	55− 91	cyt	5	−	−	
H$^+$/K$^+$ ATPase α	1−114	cyt	−	−1.10	+6	X64694
	125−145	ext	0	−	−	
	156−256	cyt	−	−0.99	−	
	267−310	ext	3	−	−	
MHV E1	1− 31	ext	2	−	−4	J02252
	42− 61	cyt	2	−	−	
	72− 88	ext	0	−	−	
	99−227	cyt	−	+0.96	−	
IBV E1	1− 6	ext	2	−	−4	X04107
	37− 57	cyt	3	−	−	
	67− 83	ext	0	−	−	
	94−224	cyt	−	+0.84	−	
CD20 receptor	1− 69	cyt	−	−0.36	+1	X07203
	81− 90	ext	0	−	−	
	101−122	cyt	5	−	−	
	132−189	ext	4	−	−	
	200−297	cyt	−	−0.91	−	
Gap-junction protein	1− 27	cyt	3	−	+4	X04325
	39− 80	ext	2	−	−	
	92−147	cyt	6	−	−	
	159−191	ext	5	−	−	
	203−283	cyt	−	−1.11	−	
MDR	1− 56	cyt	15	−	+1	M14757
	68−124	ext	1	−	−	
	136−195	cyt	7	−	−	
	207−220	ext	1	−	−	
	232−295	cyt	10	−	−	
	307−333	ext	0	−	−	
Vacuolar H$^+$ ATPase proteolipid	1− 14	ext	1	−	−3	S40059
	26− 59	cyt	3	−	−	
	71− 99	ext	1	−	−	
	111−132	cyt	2	−	−	
	144−159	ext	1	−	−	
AcChR α	25−245	ext	−	+1.37	−	P02710
	257−271	cyt	1	−	−	
	283−305	ext	1	−	−	
	317−436	cyt	−	−0.39	−	
	448−461	ext	1	−	−	

Table 1. (Continued).

Protein	Segment	Location	n_{K+R}	$D_{\text{cyt-ext}}$	Δ(N−C)	Accesion no.
STE2	1– 56	ext	1	–	−5	X03010
	68– 82	cyt	3	–	–	
	94–134	ext	2	–	–	
	146–166	cyt	3	–	–	
	178–210	ext	2	–	–	
	222–250	cyt	5	–	–	
	262–280	ext	1	–	–	
	292–431	cyt	–	0.00	–	
HMG-CoA reductase	1– 23	cyt	3	–	+1	M12705
	35– 64	ext	2	–	–	
	76– 95	cyt	2	–	–	
	107–122	ext	1	–	–	
	134–171	cyt	3	–	–	
	183–197	ext	1	–	–	
	209–259	cyt	7	–	–	
	271–326	ext	5	–	–	
Rhodopsin	1– 39	ext	3	–	−4	P02700
	59– 84	cyt	3	–	–	
	96–118	ext	0	–	–	
	130–156	cyt	3	–	–	
	168–207	ext	2	–	–	
	219–258	cyt	4	–	–	
	270–285	ext	0	–	–	
	297–349	cyt	4	–	–	
TRAM	1– 28	cyt	5	–	−1	X63679
	40– 83	ext	2	–	–	
	95–126	cyt	6	–	–	
	138–164	ext	1	–	–	
	176–200	cyt	4	–	–	
	212–221	ext	0	–	–	
	233–256	cyt	3	–	–	
	268–300	ext	3	–	–	
	312–374	cyt	15	–	–	
H^+ ATPase plasma membrane	1–120	cyt	–	−1.02	+3	J02602
	132–145	ext	0	–	–	
	157–296	cyt	–	−0.47	–	
	308–329	ext	1	–	–	
	341–656	cyt	–	−0.60	–	
	676–698	ext	4	–	–	
	710–723	cyt	1	–	–	
	735–759	ext	2	–	–	
	771–831	cyt	1	–	–	
	843–863	ext	1	–	–	
	875–919	cyt	6	–	–	

It is interesting to note that there is no significant bias in the distribution of Asp+Glu between the first cytoplasmic and periplasmic segments in the sample of prokaryotic inner-membrane proteins analyzed previously [5], 8.6% versus 9.5%, and that the 'charge difference' rule thus applies only to eukaryotic proteins.

Trp and Tyr are more prevalent in extra-cytoplasmic segments and at the ends of transmembrane stretches

As noted above, previous studies have found evidence of subtle differences in overall amino-acid composition between cytoplasmic and extra-cytoplasmic domains in transmembrane proteins, as well as between cytoplasmic and extra-cellular globular proteins [12, 45]. Our results confirm this; in particular, the frequency of (Trp+Tyr) is significantly higher in the extra-cytoplasmic segments ($P<10^{-4}$; Fig. 3). Tyr and Trp also seem to be specifically enriched near both the cytoplasmic and extra-cytoplasmic ends of transmembrane stretches (Fig. 3) in contrast to purely hydrophobic residues: the difference between the frequency of Trp+Tyr in the five residues immediately flanking the central 11 residues of the predicted transmembrane segments (8.6%) compared to their frequency in the central 11 residues of these same segments (3.9%) is highly significant ($P<10^{-3}$). This tendency of aromatic residues with polar substituents to be found in the vicinity of the lipid headgroup region has been observed previously in the photosynthetic reaction center [46], a number of β-barrel membrane proteins [4] and in single-spanning eukaryotic membrane proteins [47], and could possibly be used to predict the precise ends of transmembrane segments.

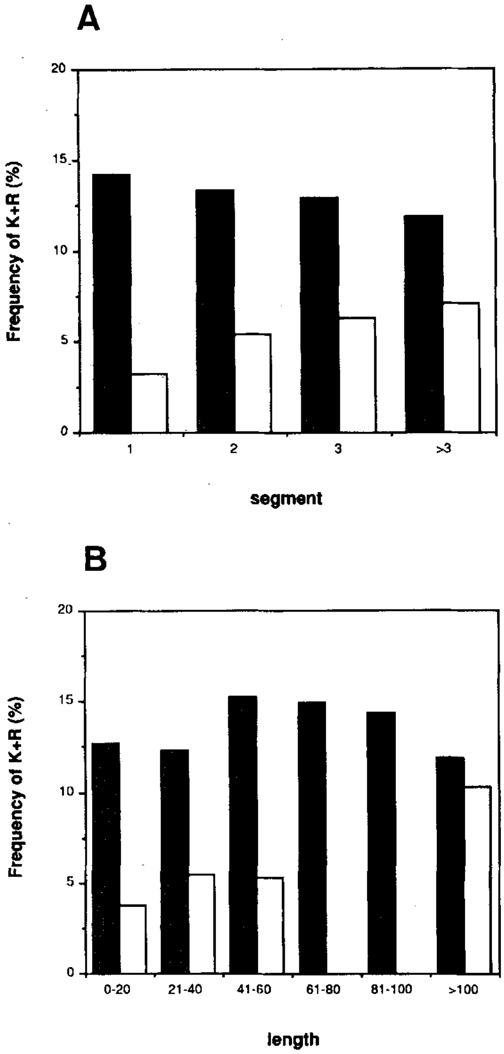

Fig. 1. Frequency of Arg+Lys relative to position and length. (A) Frequency of Arg+Lys in cytoplasmic (stippled bars) and extra-cytoplasmic polar segments (open bars) as a function of the position of the segment counting from the N-terminus (1st cytoplasmic/1st extra-cytoplasmic; 2nd cytoplasmic/2nd extra-cytoplasmic, etc.) in a sample of 16 eukaryotic multi-spanning membrane proteins. (B) Frequency of Arg+Lys in cytoplasmic (stippled bars) and extra-cytoplasmic polar segments (open bars) as a function of the length of the segment.

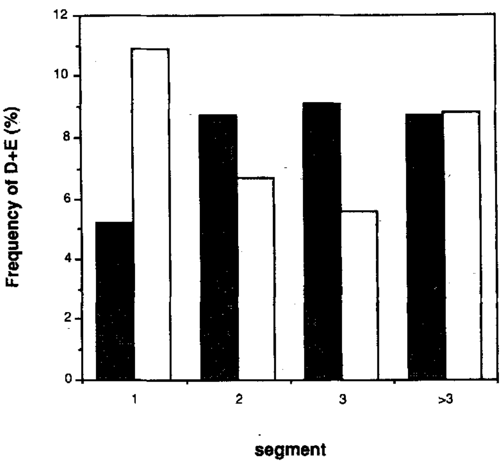

Fig. 2. Frequency of Asp+Glu in cytoplasmic (stippled bars) and extra-cytoplasmic polar segments (open bars) as a function of the position of the segment counting from the N-terminus. See Fig. 1.

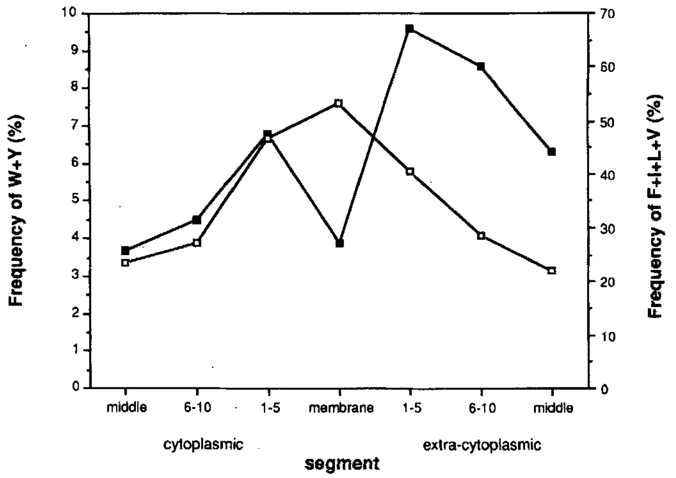

Fig. 3. Frequency of Trp+Tyr (■) and Phe+Ile+Leu+Val (□) as a function of the position relative to the membrane-spanning segments. Amino-acid frequencies were calculated for the central 11 residues in the predicted transmembrane segments (membrane), the five residues flanking this central stretch (1−5), the next five flanking residues (6−10) and the remaining residues in the polar segments (middle) for both cytoplasmic (left half) and extra-cytoplasmic (right half) segments.

Prediction of the transmembrane topology

Although a method based only on the 'positive inside' rule works very well for bacterial proteins [5], a combination of several theoretical analyses seems to be the best way to predict the topology of eukaryotic membrane proteins. For the proteins used in this study (Table 1) one (the acetylcholine-receptor α subunit) has a cleavable N-terminal signal sequence and the N-terminus of the mature protein can thus be immediately predicted (correctly) to be extra-cytoplasmic. For the remaining 15, the 'charge difference' rule predicts the correct location of the N-terminus in 13 cases (87%).

To predict the location and orientation of all transmembrane segments, we have used a combination of hydrophobicity analysis, charge bias analysis [5] and compositional-bias analysis [12]. As described in Methods, the basic strategy is, first, to identify all possible transmembrane segments from a hydrophobicity plot with suitable cutoffs, then to generate all possible topologies consistent with the candidate transmembrane segments, and finally to rank the resulting models on the basis of their degree of Arg+Lys bias (only counting polar segments shorter than 60 residues) and the overall composition of long (>60 residues) polar segments, which tends to be different for cytoplasmic and extra-cytoplasmic domains. For the 16 proteins listed in Table 1, compositional analysis predicts the correct location of 15 out of a total of 18 (83%) polar segments longer than 60 residues. Charge-bias analysis can only be applied with any confidence to 11 of the 16 proteins, since the number of long polar segments (>60 residues) is too large in the other five. For these 11 proteins, the correct topology is predicted in 10 cases (i.e. both the number of transmembrane segments as well as their orientations are correctly predicted). The one erroneous prediction (*Neurospora crassa* plasma-membrane H^+ ATPase) gives a topology (N_{out}) that is in conflict with the results from both charge difference and compositional analysis (N_{in}); if the long N-terminal polar domain is assumed to be cytoplasmic, charge-bias analysis in fact predicts the correct number of transmembrane segments and topology for the rest of the molecule.

A number of parameters that are used in this prediction scheme are known only from statistical estimates and have not been rigorously checked experimentally. Thus, hydrophobicity scales are still based on rather scant experimental data, and it is not known whether, for instance, charge-pairing between residues located one turn away from each other in a transmembrane α helix can compensate for the free-energy loss upon membrane insertion [43]. In our collection of proteins, we have found at least one experimentally mapped transmembrane segment where such charge-pairing would seem to be necessary to allow stable anchoring in the membrane (residues 123–143 of 3-hydroxy-3-methylglutaryl CoA reductase) and have therefore provisionally allowed for $i,i+3$ and $i,i+4$ charge pairing in our hydrophobicity-analysis algorithm (see Methods). Another possibility would be inter-helix charge pairing [48–51]; ultimately, this may also have to be incorporated into future prediction algorithms.

Other parameters that need to be better defined are the minimum and maximum lengths of transmembrane helices (e.g., how long can one make a hydrophobic segment before it breaks up into two neighboring transmembrane helices?); the precise number of residues on either side of an N-terminal transmembrane segment that should be included in the charge difference calculation; how the charge on the N-terminal amino group on the nascent chain should be counted (so far, we have given it full weight); and any restrictions (length, composition, etc.) on polar N-terminal tails located upstream of an N-terminal transmembrane segment with the N_{out} orientation. Finally, other features of the nascent chain such as the formation of stably folded domains have also been suggested to play a role [9].

In spite of these remaining problems, we conclude that the combination of hydrophobicity and charge-bias analysis that has previously been shown to work well for bacterial inner-membrane proteins also works for eukaryotic ones, but that it cannot be applied when there are many long, polar segments between the putative transmembrane stretches. In such cases, compositional analysis can help, although it often gives less clear-cut results. Finally, charge-difference analysis is a good predictor of the orientation of the most N-terminal transmembrane segment and can be used to constrain the possible models that go into the charge-bias analysis.

This work was supported by grants from the Swedish Natural Sciences Research Council, the Axelson-Johnson Foundation, and the Swedish Medical Research Council to GvH. We would like to thank Dr. Ylva Gavel for many valuable discussions.

REFERENCES

1. Rees, D. C., Komiya, H., Yeates, T. O., Allen, J. P. & Feher, G. (1989) *Annu. Rev. Biochem. 58*, 607–33.
2. Deisenhofer, J. & Michel, H. (1991) *Annu. Rev. Biophys. Biophys. Chem. 20*, 247–266.
3. Weiss, M. S., Kreusch, A., Schiltz, E., Nestel, U., Welte, W., Weckesser, J. & Schulz, G. E. (1991) *FEBS Lett. 280*, 379–382.
4. Cowan, S. W., Schirmer, T., Rummel, G., Steiert, M., Ghosh, R., Pauptit, R. A., Joansonius, J. N. & Rosenbusch, J. P. (1992) *Nature 358*, 727–733.
5. von Heijne, G. (1992) *J. Mol. Biol. 225*, 487–494.
6. von Heijne, G. (1986) *EMBO J. 5*, 3021–3027.
7. von Heijne, G. & Gavel, Y. (1988) *Eur. J. Biochem. 174*, 671–678.
8. Hartmann, E., Rapoport, T. A. & Lodish, H. F. (1989) *Proc. Natl Acad. Sci. USA 86*, 5786–5790.
9. Beltzer, J. P., Fiedler, K., Fuhrer, C., Geffen, I., Handschin, C., Wessels, H. P. & Spiess, M. (1991) *J. Biol. Chem. 266*, 973–978.
10. Parks, G. D. & Lamb, R. A. (1991) *Cell 64*, 777–787.
11. Sakaguchi, M., Tomiyoshi, R., Kuroiwa, T., Mihara, K. & Omura, T. (1992) *Proc. Natl Acad. Sci. USA 89*, 16–19.
12. Nakashima, H. & Nishikawa, K. (1992) *FEBS Lett. 303*, 141–146.
13. Feldheim, D., Rothblatt, J. & Schekman, R. (1992) *Mol. Cell. Biol. 12*, 3288–3296.
14. Deshaies, R. J. & Schekman, R. (1990) *Mol. Cell. Biol. 10*, 6024–6035.
15. Krajcsi, P., Tollefson, A. E., Anderson, C. W., Stewart, A. R., Carlin, C. R. & Wold, W. S. M. (1992) *Virology 187*, 131–144.
16. Smolka, A., Alverson, L., Fritz, R., Swiger, K. & Swiger, R. (1991) *Biochem. Biophys. Res. Commun. 180*, 1356–1364.
17. Bamberg, K., Mercier, F., Reuben, M. A., Kobayashi, Y., Munson, K. B. & Sachs, G. (1992) *Biochim. Biophys. Acta 1131*, 69–77.
18. Munson, K. B., Gutierrez, C., Balaji, V. N., Ramnarayan, K. & Sachs, G. (1991) *J. Biol. Chem. 266*, 18976–18988.
19. Smolka, A. & Swiger, K. M. (1992) *Biochim. Biophys. Acta 1108*, 75–85.
20. Mayer, T., Tamura, T., Falk, M. & Niemann, H. (1988) *J. Biol. Chem. 263*, 14956–14963.
21. Machamer, C. E. & Rose, J. K. (1987) *J. Cell. Biol. 105*, 1205–1214.
22. Einfeld, D. A., Brown, J. P., Valentine, M. A., Clark, E. A. & Ledbetter, J. A. (1988) *EMBO J. 7*, 711–7.
23. Evans, W. H. & Rahman, S. (1989) *Biochem. Soc. Trans. 17*, 983–985.
24. Milks, L. C., Kumar, N. M., Houghten, R., Unwin, N. & Gilula, N. B. (1988) *EMBO J. 7*, 2967–2975.
25. Yeager, M. & Gilula, N. B. (1992) *J. Mol. Biol. 223*, 929–948.
26. Juranka, P. F., Zastawny, R. L. & Ling, V. (1989) *FASEB J. 3*, 2583–2592.
27. Georges, E., Bradley, G., Gariepy, J. & Ling, V. (1990) *Proc. Natl Acad. Sci. USA 87*, 152–156.
28. Yoshimura, A., Kuwazuru, Y., Sumizawa, T., Ichikawa, M., Ikeda, S., Uda, T. & Akiyama, S. (1989) *J. Biol. Chem. 264*, 16282–16291.
29. Zhang, J. T. & Ling, V. (1991) *J. Biol. Chem. 266*, 18224–18232.
30. Finbow, M. E., Eliopoulos, E. E., Jackson, P. J., Keen, J. N., Meagher, L., Thompson, P., Jones, P. & Findlay, J. B. C. (1992) *Protein Eng. 5*, 7–15.
31. Chavez, R. A. & Hall, Z. W. (1991) *J. Biol. Chem. 266*, 15532–15538.
32. Chavez, R. A. & Hall, Z. W. (1992) *J. Cell. Biol. 116*, 385–393.
33. DiPaola, M., Czajkowski, C. & Karlin, A. (1989) *J. Biol. Chem. 264*, 15457–15463.
34. Reneke, J. E., Blumer, K. J., Courchesne, W. E. & Thorner, J. (1988) *Cell 55*, 221–234.
35. Cartwright, C. P. & Tipper, D. J. (1991) *Mol. Cell. Biol. 11*, 2620–2628.
36. Roitelman, J., Olender, E. H., Barnun, S., Dunn, W. A. & Simoni, R. D. (1992) *J. Cell. Biol. 117*, 959–973.
37. Olender, E. H. & Simoni, R. D. (1992) *J. Biol. Chem. 267*, 4223–4235.
38. Sengstag, C., Stirling, C., Schekman, R. & Rine, J. (1990) *Mol. Cell. Biol. 10*, 672–680.
39. Findlay, J. B. & Pappin, D. J. (1986) *Biochem. J. 238*, 625–642.
40. Görlich, D., Hartmann, E., Prehn, S. & Rapoport, T. A. (1992) *Nature 357*, 47–52.

41. Scarborough, G. A. (1992) *Mol. Cell. Biochem. 114*, 49–56.
42. Engelman, D. M., Steitz, T. A. & Goldman, A. (1986) *Annu. Rev. Biophys. Biophys. Chem. 15*, 321–353.
43. Honig, B. H. & Hubbell, W. L. (1984) *Proc. Natl Acad. Sci. USA 81*, 5412–5416.
44. Andersson, H. & von Heijne, G. (1993) *EMBO J. 12*, 683–691.
45. Nishikawa, K., Kubota, Y. & Ooi, T. (1983) *J. Biochem. (Tokyo) 94*, 997–1007.
46. Schiffer, M., Chang, C. H. & Stevens, F. J. (1992) *Protein Eng. 5*, 213–214.
47. Landolt-Marticorena, C., Williams, K. A., Deber, C. M. & Reithmeier, R. A. F. (1993) *J. Mol. Biol.*, in the press.
48. Cosson, P., Lankford, S. P., Bonifacino, J. S. & Klausner, R. D. (1991) *Nature 351*, 414–416.
49. Lee, J.-I., Hwang, P. P., Hansen, C. & Wilson, T. H. (1992) *J. Biol. Chem. 267*, 20758–20764.
50. Sahin-Tóth, M., Dunten, R. L., Gonzales, A. & Kaback, H. R. (1992) *Proc. Natl Acad. Sci. USA 89*, 10547–10551.
51. Howitt, S. M. & Cox, G. B. (1992) *Proc. Natl Acad. Sci. USA 89*, 9799–9803.