Protein Science (1998), 7:1029–1038. Cambridge University Press. Printed in the U.S.A.

Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms

ERIK WALLIN AND GUNNAR VON HEIJNE

Department of Biochemistry, Stockholm University, S-106 91 Stockholm, Sweden

(RECEIVED September 30, 1997; ACCEPTED January 13, 1998)

Abstract

We have carried out detailed statistical analyses of integral membrane proteins of the helix-bundle class from eubacterial, archaean, and eukaryotic organisms for which genome-wide sequence data are available. Twenty to 30% of all ORFs are predicted to encode membrane proteins, with the larger genomes containing a higher fraction than the smaller ones. Although there is a general tendency that proteins with a smaller number of transmembrane segments are more prevalent than those with many, uni-cellular organisms appear to prefer proteins with 6 and 12 transmembrane segments, whereas *Caenorhabditis elegans* and *Homo sapiens* have a slight preference for proteins with seven transmembrane segments. In all organisms, there is a tendency that membrane proteins either have many transmembrane segments with short connecting loops or few transmembrane segments with large extra-membraneous domains. Membrane proteins from all organisms studied, except possibly the archaeon *Methanococcus jannaschii*, follow the so-called "positive-inside" rule; i.e., they tend to have a higher frequency of positively charged residues in cytoplasmic than in extracytoplasmic segments.

Keywords: genomics; membrane protein; topology

The topology of a membrane protein, i.e., a specification of its transmembrane segments and its overall orientation in the membrane, is a fundamental structural characteristic. Over the past few years, experimental as well as statistical studies of so-called helix bundle membrane proteins from many different membrane systems have pointed to charged residues flanking the hydrophobic transmembrane segments as the predominant topological determinants (von Heijne, 1997), although other properties such as the length of a hydrophobic segment and the folding kinetics of extramembraneous domains also contribute (Denzer et al., 1995; Spiess, 1995; Wahlberg & Spiess, 1997).

This so-called positive-inside rule was first established for bacterial inner membrane proteins (mainly from *Escherichia coli*) by statistical studies (von Heijne, 1986), and was subsequently found to hold also for eukaryotic plasma membrane proteins (von Heijne & Gavel, 1988), thylakoid membrane proteins (Gavel et al., 1991), and mitochondrial inner membrane proteins encoded in the organellar genome (Gavel & von Heijne, 1992). Although the tendency of positively charged residues to be enriched in non-translocated parts of a protein is thus common to proteins from all these membrane systems, certain differences have also been noted. In particular, although the translocated parts of bacterial inner membrane proteins are impoverished in positively charged residues compared to globular periplasmic proteins (von Heijne, 1997), a reduced frequency of Arg and Lys is not seen in translocated parts of eukaryotic plasma membrane proteins (Wallin & von Heijne, 1995). Furthermore, while the most N-terminal transmembrane segment in eukaryotic plasma membrane proteins is characterized by a "negative-outside" distribution of Asp and Glu in addition to the "positive-inside" distribution of Arg and Lys (Hartmann et al., 1989; Sipos & von Heijne, 1993; Wallin & von Heijne, 1995), no such bias has been detected in the bacterial proteins. Experimentally, the importance of charged residues for membrane protein topology has been established for *E. coli* and mammalian plasma membrane proteins (Spiess, 1995; von Heijne, 1997).

The availability of complete or partial genome sequences for a number of organisms from the eubacterial, archaean, and eukaryotic domains now makes possible much more detailed studies of correlations between membrane protein topology and amino acid distributions. Here, we report a statistical analysis of multi-spanning membrane proteins from *E. coli*, *Haemophilus influenzae*, and *Helicobacter pylori* (all Gram-negative eubacteria), *Synechocystis* sp. (a Gram-negative cyanobacterium), *Bacillus subtilis* and *Clostridium acetobutylicum* (Gram-positive eubacteria), *Mycoplasma genitalium* and *Mycoplasma pneumoniae* (parasitic Gram-positive eubacteria), *Methanococcus jannaschii*, *Methanobacterium thermoautotrophicum*, and *Archaeglobus fulgidus* (archaea), *Saccharomyces cerevisiae* (a fungus), *Caenorhabditis elegans* (a nematode), and *Homo sapiens* (a mammal). In keeping with the earlier studies on much smaller samples of multi-species composition, integral membrane proteins from all organisms show a strong correlation between topology and the distribution of positively but not nega-

Reprint requests to: Gunnar von Heijne, Department of Biochemistry, Stockholm University, S-106 91 Stockholm, Sweden; e-mail: gunnar@biokemi.su.se.

tively charged residues; the only possible exception being *M. jannaschii*, which has an apparently much weaker bias in the distribution of Arg and Lys.

We have also characterized the predicted membrane proteins in the various organisms in terms of the overall frequency of membrane-protein encoding ORFs in each genome and in terms of number of transmembrane segments. In general, we find that 20–30% of all ORFs encode integral membrane proteins, and that there is an apparent preference for 6 and 12 transmembrane segments among the unicellular organisms and a weak preference for 7 transmembrane segments in *C. elegans* and *H. sapiens*. Interestingly, membrane proteins seem to come in two basic varieties: those with many transmembrane segments and short connecting loops, and those with few transmembrane segments and large extramembraneous domains; this pattern is particularly prominent among eubacterial and archaean organisms.

Results

20–30% of all ORFs encode integral membrane proteins

To get an idea of the total number of integral membrane proteins in each organism, we carried out hydrophobicity analysis predictions with different selection criteria. Because cleavable signal peptides often score as transmembrane segments (Nielsen et al., 1997), we imposed the requirement that there should be a minimum of two predicted transmembrane segments. We thus counted all ORFs encoding a protein with at least two segments predicted as "putative" by the TOPPRED algorithm (von Heijne, 1992; Claros & von Heijne, 1994), with at least one segment predicted as "certain" and one as "putative," and with at least two segments predicted as "certain" (Table 1). The first number is certainly an overestimate since "putative" segments are found quite frequently also in globular proteins (von Heijne, 1992), while the last number may be a slight underestimate. On balance, this analysis suggests that some 20–30% of all ORFs encode integral membrane proteins [not including β-barrel proteins of the porin type (Cowan & Rosen-

Table 1. *Incidence of predicted membrane proteins*[a]

Organism	Total ORFs	Set 1 (%)	Set 2 (%)	Set 3 (%)
B. subtilis	2,501	33	29	23
C. acetobutylicum	4,018	28	25	18
E. coli	4,285	40	34	24
H. influenzae	1,680	30	26	18
H. pylori	1,590	32	27	19
M. genitalium	468	29	25	18
M. pneumoniae	300	29	23	16
Synechocystis sp.	3,168	41	35	24
M. thermoautotrophicum	1,998	30	25	18
M. jannaschii	1,735	23	20	14
A. fulgidus	2,437	30	26	19
S. cerevisiae	6,218	40	34	23
C. elegans	13,201	46	41	30
H. sapiens	10,442	43	37	26

[a]Set 1: at least two "putative" or "certain" transmembrane segments. Set 2: at least one "certain" and one "putative" or "certain" transmembrane segment. Set 3: at least two "certain" transmembrane segments.

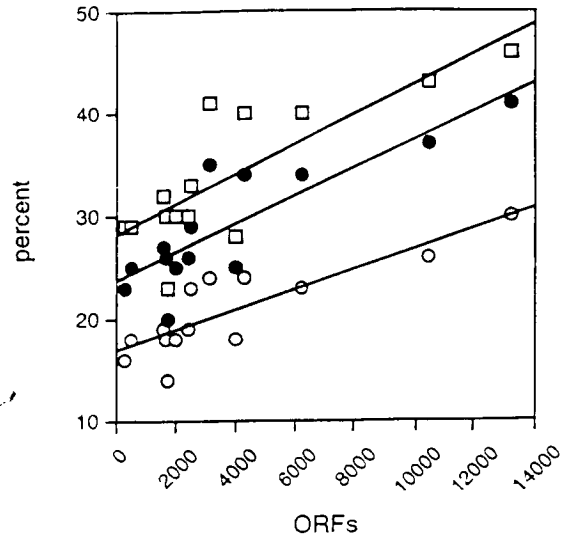


Fig. 1. Larger genomes contain a larger fraction of ORFs encoding membrane proteins. Three different estimates of the number of ORFs are shown as detailed in Table 1. Set 1: white squares. Set 2: black circles. Set 3: white circles. Linear fits to the data points are also shown. Note that only ~10,000 of the estimated 70,000 *H. sapiens* ORFs have been analyzed.

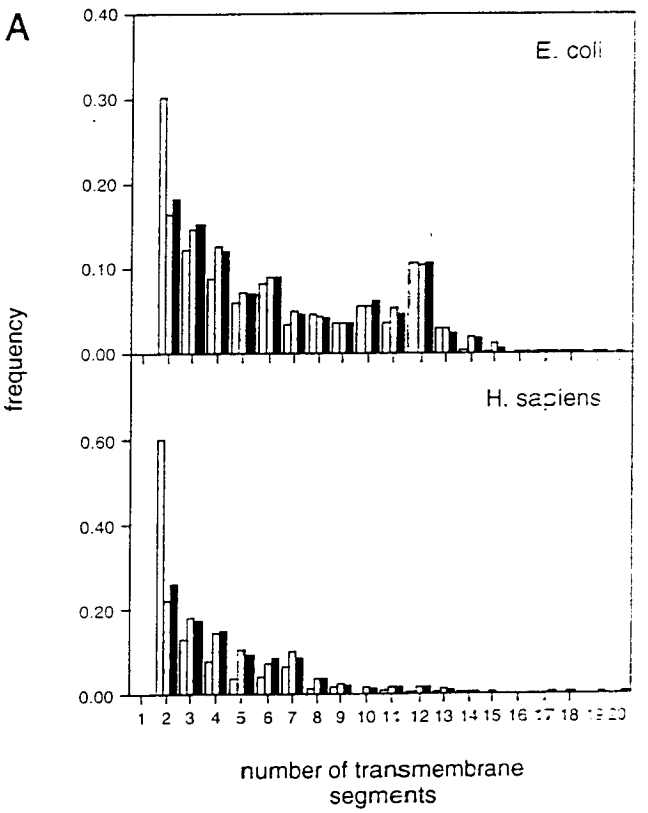


Fig. 2. Fraction of membrane proteins with different numbers of predicted transmembrane segments. **A:** Results for *E. coli* and *H. sapiens* using three different prediction schemes. White bars: proteins with only "certain" and no "putative" transmembrane segments. Gray bars: both "certain" and "putative" transmembrane segments included. Black bars: top-ranking TOPPRED predictions. **B:** Top-ranking TOPPRED predictions (proteins in Set 3, Table 1) for all organisms. Six and 12 transmembrane segments are indicated by dashed lines. (*Figure continues on facing page.*)

busch, 1994)]. Interestingly, there is a fairly good correlation between the fraction of the ORFs that encode membrane proteins and the total number of ORFs in the genome (Fig. 1), suggesting that more complex organisms tend to need a disproportionately larger complement of membrane proteins.

12 TM proteins are over-represented in most uni-cellular organisms and 7 TM proteins in C. elegans and H. sapiens

Because current topology prediction methods are quite reliable (von Heijne, 1992; Jones et al., 1994; Rost et al., 1996; Persson &

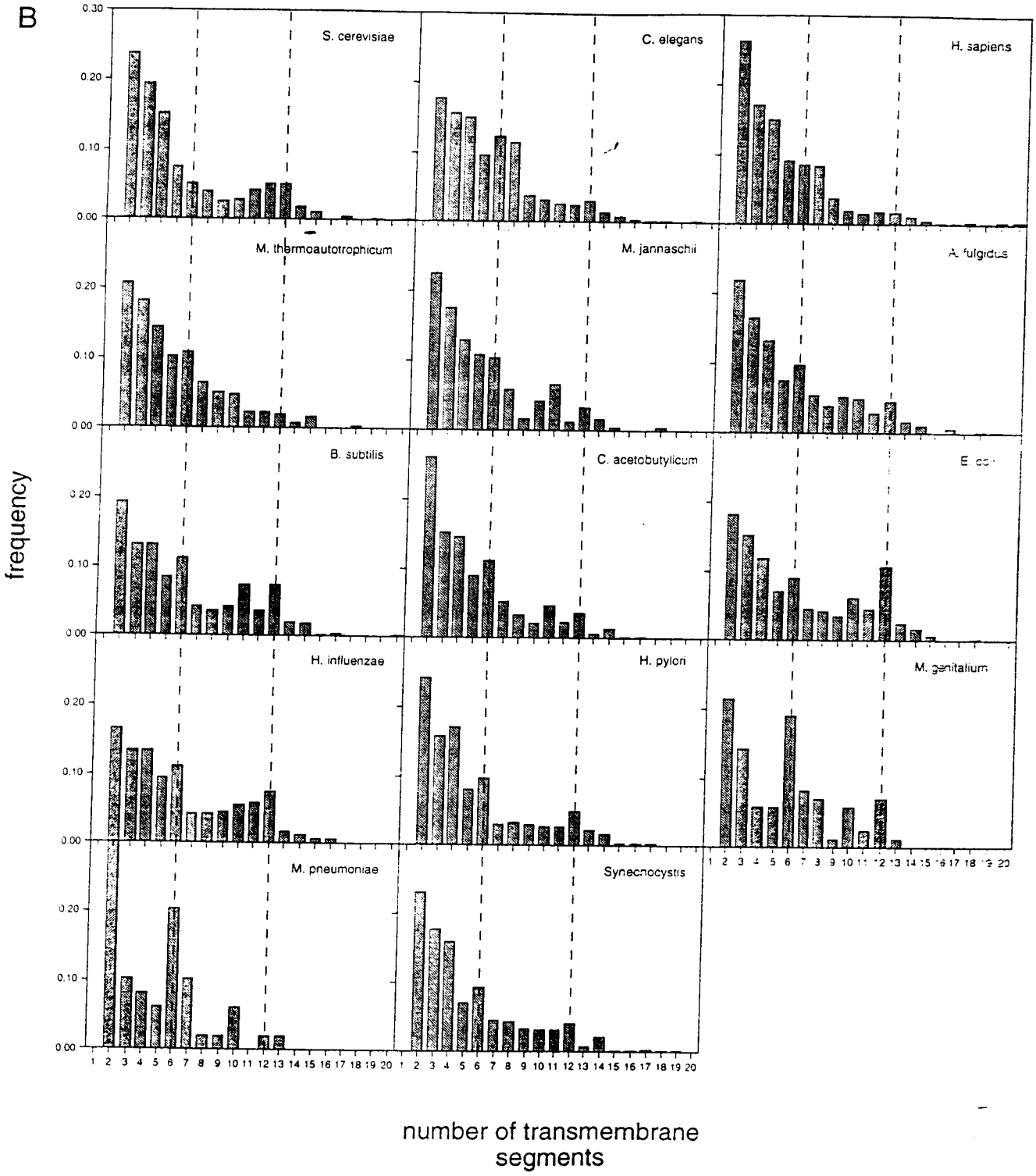


Fig. 2. *Continued.*

Argos, 1997), we collected statistics on the probable number of transmembrane (TM) segments in the different ORFs. Only proteins that have two or more segments predicted as "certain" transmembrane segments by TOPPRED (set 3, Table 1) were included. The number of transmembrane segments was predicted in three ways: first, by ignoring all proteins with one or more "putative" segments (i.e., only including proteins where all TMs were predicted as "certain"); second, by counting both the "certain" and the "putative" segments; and third, by taking the top-ranking topology from the full TOPPRED prediction for each protein (including both "certain" and "putative" segments). The first prediction should be fairly reliable, but excludes a large number of proteins from the analysis; the second prediction tends to overestimate the number of transmembrane segments; and the third prediction should again be fairly reliable, at least for those organisms where the positive-inside rule holds well (see below).

As seen in Figure 2A, the results from the three different prediction schemes are remarkably similar. For all organisms, there is a clear trend that proteins with a smaller number of transmembrane segments are more common, and very few proteins have more than 12 transmembrane segments (Fig. 2B). The rapid fall off in the number of proteins with increasing numbers of transmembrane segments has been noted recently (Arkin et al., 1997; Gerstein, 1997). To study this further, we collected statistics on the relation between the number of transmembrane segments and overall protein length. Representative results are shown in Figure 3. There is an interesting clustering clearly visible in these diagrams: there are many membrane proteins with a large number of transmembrane segments and short connecting loops, and many with only a couple of transmembrane segments and large extra-membraneous domains, but only few with multiple transmembrane segments and large extra-membraneous domains. This pattern is particularly strong among the eubacteria and archaea, but is also detectable in the eukaryotes (*S. cerevisiae*, *C. elegans*, and *H. sapiens*). For the multi-spanning proteins, the overall length increases by about 36 residues for each new transmembrane segment. The "elementary building block" in these proteins is thus a 20–25 residues long, hydrophobic transmembrane α-helix plus a 10–15 residues long loop.

As seen in Figures 2 and 3, most of the eubacterial and archaean organisms together with *S. cerevisiae* have a local peak in the distribution at 12 predicted transmembrane segments and an overall length of ~375–475 residues (~550 residues in *S. cerevisiae*), while *C. elegans* and *H. sapiens* have an apparent local peak at seven predicted transmembrane segments and overall length of ~300–350 residues. Most of the eubacterial and archaean organisms also have a local peak at six transmembrane segments (~225–275 residues); this peak is particularly strong in the two *Mycoplasma* species. Although about two-thirds of the *E. coli* proteins predicted by TOPPRED to belong to these classes are listed as "hypothetical" or have no annotation, 22 of the 29 annotated proteins with 6 transmembrane segments from *E. coli* are transporters for small solutes such as water, sulfate, phosphate, formate, and putrescine/spermidine, and 36 of the 43 annotated proteins with 12 transmembrane segments are amino acid transporters, sugar transporters, or belong to the family of ABC transporters (Beckmann et al., 1997). Not surprisingly, 113 of the 139 annotated *H. sapiens* proteins with 7 predicted transmembrane segments are listed as belonging to the 7TM family of G-protein coupled receptors.

Analysis of amino acid biases

Although previous statistical studies of amino acid biases in membrane proteins have been based on proteins with experimentally determined topologies, this obviously is not possible when genome data are used. Instead, we have sought to extract sequences that

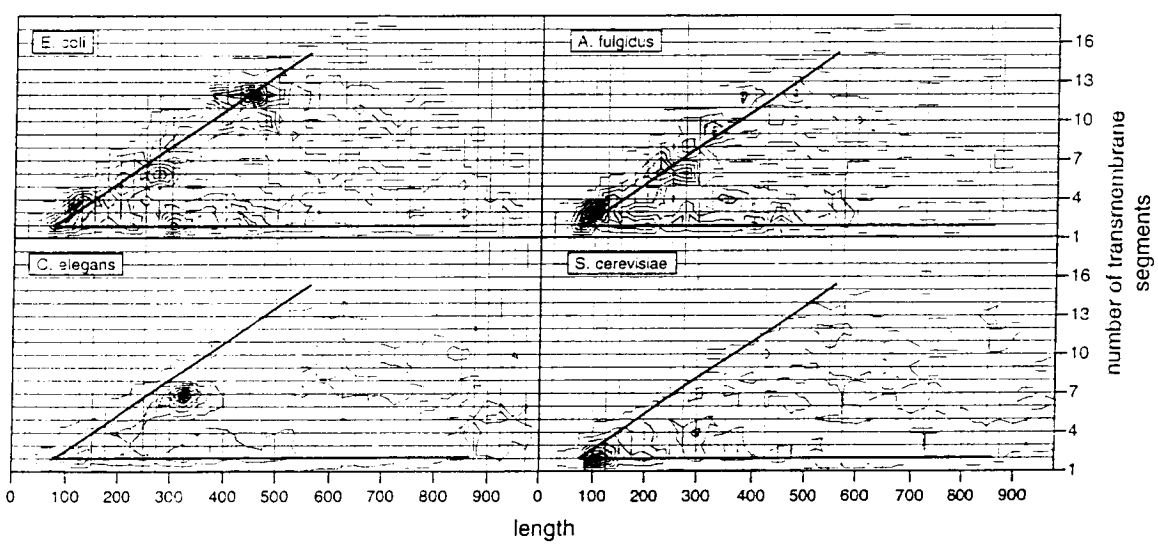


Fig. 3. Contour plots showing the frequency of proteins with a given number of predicted transmembrane segments and given overall length (in bins of 25 residues) for proteins in four representative organisms. The data sets are the same as in Figure 2B, i.e., the full TOPPRED predictions for Set 3 in Table 1. Black denotes the highest frequency value in each panel, and progressively lighter shades of gray indicate lower frequencies. The two full lines in each panel indicate the two membrane protein varieties discussed in the text. For *E. coli*, each contour level represents an increase in frequency by three proteins, and the maximum number of proteins in any one data point (black) is 23 (12 transmembrane segments, 425–450 residues length). The corresponding values for the other organisms are: 2 and 15 proteins (*A. fulgidus*), 15 and 122 proteins (*C. elegans*), and 7 and 58 proteins (*S. cerevisiae*).

(a) have a clear-cut pattern of very hydrophobic candidate transmembrane segments to ensure that the number of transmembrane segments can be predicted with high confidence, and (b) have a large number (seven or more) of predicted transmembrane segments as the method we use to characterize amino acid biases works only when the number of transmembrane segments is relatively high. In addition, we have restricted our analysis of amino acid biases to proteins in which all loops and N- and C-terminal tails are no longer than 60 residues, because it has been previously found that long loops and tails tend to have a less biased amino acid composition (von Heijne, 1986; von Heijne & Gavel, 1988; Wallin & von Heijne, 1995).

For each organism, hydrophobicity plots of all the ORFs noted in the respective sequence database were made using the TOPPRED program, and only those sequences that had seven or more "certain" candidate transmembrane segments and no "putative" candidate segments were selected (see Methods). As explained above, we also required that all loops and the N- and C-terminal tails should be ≤60 residues long. Between 0.3 and 3.8% of the sequences in the various genomes survived this selection step (Table 2). This corresponds to between 1 and 10% of all predicted membrane proteins in the different organisms (see above). The two *Mycoplasma* species were excluded from further analysis, because too few proteins remained after the selection step. Altogether, 770 proteins were included in the analysis. To estimate the number of nonhomologous proteins in this collection, the method of Hobohom et al. (Hobohom et al., 1992) was used to select a subset of sequences where no pair was more than 30% identical over a stretch of 80 residues or more; this reduced set contained 263 nonhomologous sequences. The results are thus not dominated by a small number of highly populated families.

One problem with hydrophobicity-based prediction methods is that two closely spaced transmembrane segments are sometimes predicted as one very long segment. To control for this, we repeated the selection but now using a shorter window (17 rather than 21 residues) in the hydrophobicity analysis, and kept only those sequences that had seven or more "certain" and no "putative" candidate transmembrane segments in both selections. In general, about 40% of the sequences in the first selection survived this second round (Table 2).

For each protein k and for each kind of amino acid i we calculated the absolute difference in frequency between the two sides of the predicted topology, $\Delta f_{i,k} = |f^e_{i,k} - f^o_{i,k}|$, where e refers to even-numbered and o to odd-numbered loops (Fig. 4A). For the same protein, we then generated all possible permutations of the loops between the two sides of the structure and calculated the mean of the absolute frequency difference, $\langle \Delta f_{i,k} \rangle$, over this set. Finally, $(\Delta f_{i,k} - \langle \Delta f_{i,k} \rangle)$ was averaged over all proteins from a given organism and taken as a measure of the deviation from a random distribution for the given residue.

We also used a crude binomial test to assess the statistical significance of the observed deviations, where we counted, for each amino acid, the number of proteins from the given organism that had a deviation larger than the median $M_{i,k}$ of the $\Delta f_{i,k}$ values calculated for the permuted set of sequences, and then calculated the probability that the observed or a larger number of proteins would have $\Delta f_{i,k}$ values larger than $M_{i,k}$, as described in Methods.

Table 2. *Number of membrane proteins included in the amino acid bias analysis*[a]

Organism	First selection	Second selection
B. subtilis	97	42
C. acetobutylicum	63	28
E. coli	152	62
H. influenzae	44	16
Synechocystis sp.	40	11
H. pylori	28	11
M. thermoautotrophicum	26	11
M. jannaschii	29	12
A. fulgidus	61	24
S. cerevisiae	18	8
C. elegans	169	69
H. sapiens	43	28

[a]The first selection is based on the requirement that the proteins must have at least seven "certain" and no "putative" transmembrane segments and no loops or tails >60 residues as predicted by TOPPRED with a full window length of 21 residues; in the second selection, this requirement must be fulfilled both for window lengths of 21 and 17 residues. See http://www.biokemi.su.se/~gvh/genome-tm-analysis.html for a complete listing of these proteins.

Eubacterial membrane proteins follow the positive-inside rule

The charge bias for the eubacterial organisms using the less stringent selection criterion (first selection in Table 2) are shown in Figure 4B (two bottom rows); very similar results were obtained with the more stringently selected second set of sequences (data not shown). The patterns are very similar and (Arg − Lys) shows a significantly higher bias ($p < 10^{-2}$) in the real proteins compared to the randomized controls except for *H. pylori* (this is probably only a result of the small number of sequences because Lys and (Arg + Lys) stand out against all other residues also in this case). The bias for (Arg + Lys) is in all cases higher than for either of these two residues taken separately, showing that Arg and Lys tend to occur on the same side of the proteins. This is consistent with previous statistical studies as well as with the well-documented effect of Arg and Lys residues on the topology of *E. coli* inner membrane proteins (von Heijne, 1997). Interestingly, Lys tends to be more biased than Arg (except for *E. coli*). As seen in Figure 5 (bottom panel), there is not much difference between the (Arg + Lys) biases calculated for the full sets of sequences (white bars) and those remaining after selecting with the two different window lengths (gray bars); the latter biases tend to be slightly stronger. The result for *E. coli* thus suggests that our selection criteria are valid, and the results for the other species suggest that the "rules" for membrane protein topology are essentially the same in all eubacteria.

Eukaryotic membrane proteins follow the positive-inside rule

Results for *S. cerevisiae*, *C. elegans*, and *H. sapiens* are shown in Figure 4B (top row). Again, the distribution of (Arg − Lys) shows a significant bias for *C. elegans* and *H. sapiens* and appears skewed for *S. cerevisiae*, although the latter is not significant on the 1% level by the binomial test (14 out of 18 proteins have a larger bias than in the randomized controls, $p = 0.02$), possibly as a result of the rather limited number of proteins in the sample. There is also a slight but significant bias of (Asp + Glu) in *C. elegans* and

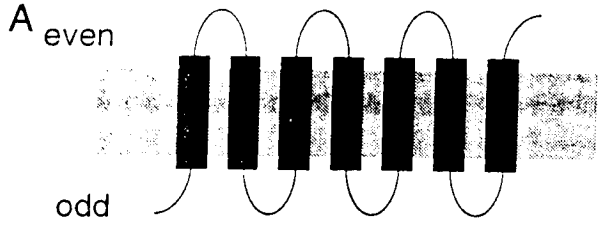
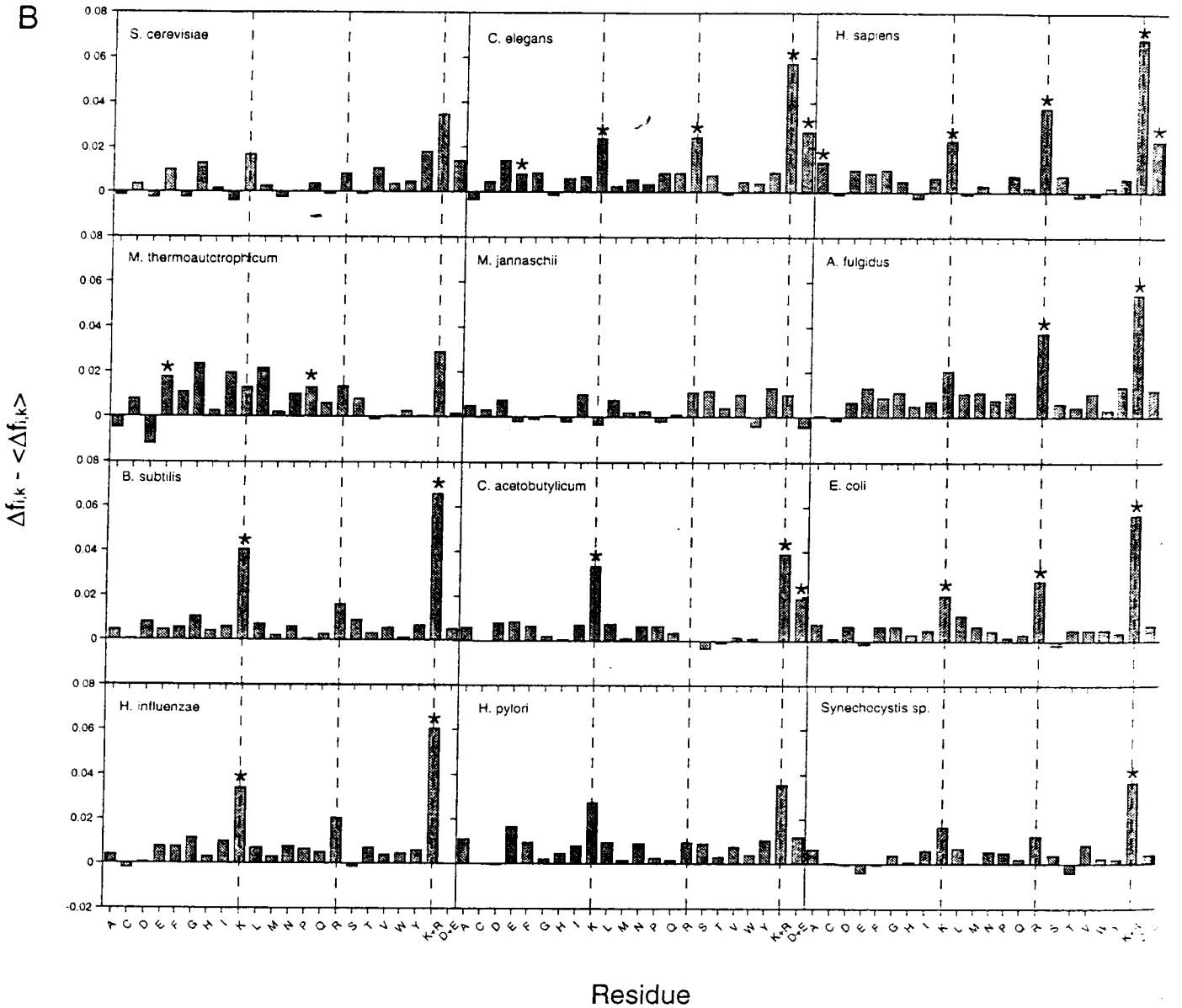


Fig. 4. Amino acid biases for different organisms. **A:** All selected proteins from the different organisms have seven or more predicted transmembrane segments and all loops and tails are ≤60 residues long. Even- and odd-numbered loops and tails are on opposite sides of the membrane. **B:** Amino acid biases for proteins in the first selection (see Table 2). Biases that are significant on the 1% level as assessed by a binomial test (see Methods) are indicated by *. Data for Lys, Arg, and (Arg + Lys) are indicated by dashed lines.



H. sapiens. This bias is opposite to the (Arg + Lys) bias, i.e., if the side with the higher frequency of (Arg + Lys) is take to be "in" the N-terminal segment with the higher number of Asp + Glu tends to be "out," as seen from the stronger bias in the net charge (Arg + Lys-Asp-Glu) than the total charge (Arg + Lys + Asp + Glu) (Fig. 5, top panel).

Charge biases in archaean membrane proteins

The amino acid bias profiles for the three archaean organisms are shown in Figure 4B (second row). While *A. fulgidus* has a strong bias for (Arg + Lys), *M. thermoautotrophicum* has a weak bias that is not statistically significant using the binomial test, and there is at best a very small bias for the *M. jannaschii* proteins. The (Arg + Lys) bias for *M. jannaschii* is, however, markedly stronger when proteins are selected with two different window lengths (Fig. 5, bottom panel), suggesting that prediction errors may at least in part explain the very weak bias seen in Figure 4B. In all three cases, the bias for (Arg + Lys) is markedly larger than for (Asp + Glu). Interestingly, the net charge bias is quite strong for all three organisms (Fig. 5, top panel).

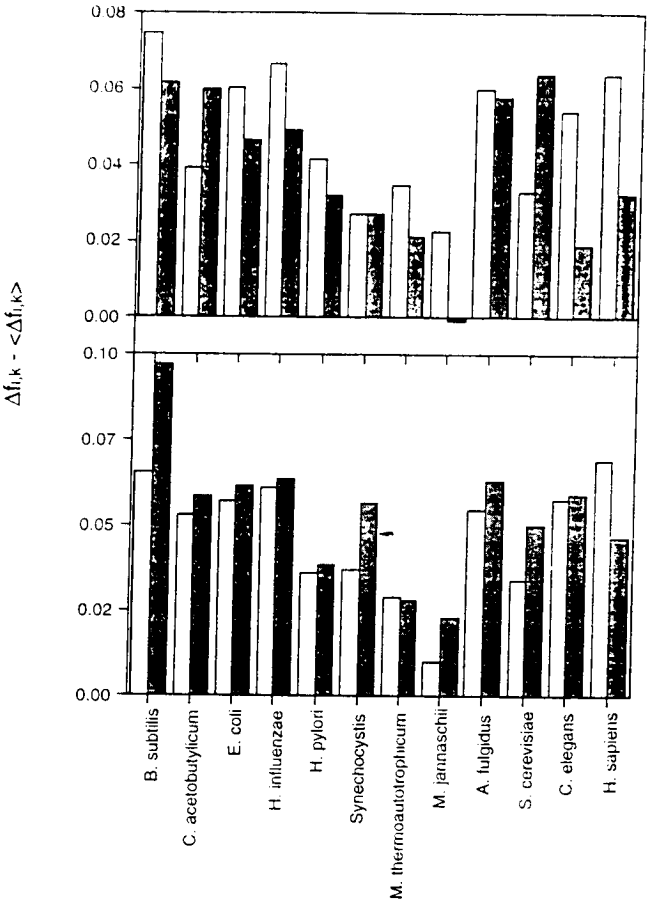


Fig. 5. Top panel: Net charge bias (Arg + Lys-Asp-Glu; white bars) and total charge bias (Arg + Lys + Asp + Glu; gray bars) for different organisms. Bottom panel: (Arg + Lys) bias for proteins in the first (white bars) and second (gray bars) selections (see Table 2).

Charge distribution in proteins with known topology

Analysis of the absolute bias in amino acid distributions suffers from the drawback that a correlation between amino acid distributions and the inside/outside orientation of the proteins cannot be made. To get an idea whether the more positively charged side correlates with an inside (i.e., cytoplasmic) localization, we sought to find multi-spanning membrane proteins with an experimentally determined topology in at least one organism that are present in most or all of the other genomes analyzed above. The only such protein we have found so far is SecY/Sec61α, the central component of the preprotein translocase (Rapoport et al., 1996). SecY sequences are known from all the organisms in our study except *H. sapiens* (we used the rat Sec61α sequence instead). The *E. coli* SecY and *S. cerevisiae* Sec61p are known to have 10 transmembrane segments and an N_{cyt}-C_{cyt} topology (Akiyama & Ito, 1987; Wilkinson et al., 1996), and sequence alignments suggest that the other SecY/Sec61α proteins also have 10 transmembrane segments. As shown in Table 3, there is a strong accumulation of Arg and Lys in the predicted cytoplasmic segments of all these proteins. In agreement with the results reported above, the difference between the cytoplasmic and extra-cytoplasmic sides is more marked in the eubacteria and archaea than in the eukaryotes, mainly because the frequency in the extra-cytoplasmic loops is higher in the latter. Interestingly, there is a clear accumulation of (Arg − Lys) in the cytoplasmic segments even for *M. jannaschii* SecY, suggesting that the positive-inside rule may in fact hold also for this archaeon.

Table 3. *Charge biases in SecY/Sec61α proteins from different organisms*[a]

Organism	$n_{in} - n_{out}$	f_{in}	f_{out}	$f_{in} - f_{out}$
B. subtilis	21	0.18	0.06	0.12
C. acetobutylicum	18	0.18	0.08	0.10
E. coli	25	0.20	0.07	0.13
H. influenzae	25	0.20	0.06	0.14
H. pylori	21	0.19	0.05	0.14
M. genitalium	24	0.16	0.04	0.11
M. pneumoniae	28	0.17	0.03	0.14
Synechocystis sp.	25	0.19	0.05	0.14
M. thermoautotrophicum	23	0.18	0.05	0.13
M. jannaschii	26	0.20	0.04	0.16
A. fulgidus	21	0.17	0.04	0.13
S. cerevisiae	10	0.17	0.12	0.05
C. elegans	15	0.18	0.09	0.09
R. norvegicus	14	0.17	0.09	0.09

[a]The net difference in the number of Arg + Lys residues between the cytoplasmic (in) and extra-cytoplasmic (out) segments ($n_{in} - n_{out}$), the corresponding frequencies (f_{in}, f_{out}) and the frequency difference are given. The identification of the transmembrane segments was based on a multiple sequence alignment (see Methods).

Because the bias in (Arg + Lys) reported for *H. pylori* in Figure 4B was not statistically significant as judged by the binomial test, we also analyzed a couple of *H. pylori* proteins (HP0299, HP1091, and HP1077) with high similarity (35–45% identity) to other eubacterial proteins with experimentally determined topologies (OPPC_SALTY, KGTP_ECOLI, and HOXN_ALCEU; see Cserzö et al., 1997, for references). In all cases the *H. pylori* proteins had most of their Arg and Lys residues in the cytoplasmic segments, and the degree of bias was about equal to that in the corresponding proteins from the other eubacterial organisms (Table 4).

Discussion

Statistical analysis of helix bundle integral membrane proteins selected from genome-wide sequence data representing the three basic domains of living organisms, i.e., archaea, eubacteria, and

Table 4. *Charge biases in H. pylori proteins (TIGR code) and in homologues (SwissProt code) of known topology*[a]

Protein/homologue	$n_{in} - n_{out}$	f_{in}	f_{out}	$f_{in} - f_{out}$
HP0299/OPPC_SALTY	9/17	0.19/0.18	0.08/0.04	0.11/0.14
HP1091/KGTP_ECOLI	21/25	0.21/0.22	0.08/0.08	0.13/0.14
HP1077/HOXN_ALCEU	14/21	0.15/0.18	0.10/0.09	0.05/0.09

[a]The net difference in the number of Arg + Lys residues between the cytoplasmic (in) and extra-cytoplasmic (out) segments ($n_{in} - n_{out}$), the corresponding frequencies (f_{in}, f_{out}) and the frequency difference are given.

eukaryota, have been carried out in an attempt to assess the validity of the positive-inside rule and to look for further correlations between amino acid distributions and topology. We have also made a rough estimate of the total number of membrane proteins in each organism and the relative incidence of proteins with different numbers of transmembrane segments.

As shown in Table 1, membrane proteins of the helix bundle class generally account for 20–30% of the ORFs in the various genomes, with the larger genomes having somewhat higher fractions than the smaller ones. This value was arrived at by calculating hydrophobicity profiles according to the TOPPRED algorithm (von Heijne, 1992; Claros & von Heijne, 1994) and requiring that at least two predicted transmembrane segments are present. This requirement was chosen to avoid as much as possible the erroneous inclusion of secreted proteins [signal peptides often score as transmembrane segments (Nielsen et al., 1997)], and the only membrane proteins that are completely missed in this way are the single-spanning signal-anchor and tail-anchored (Kutay et al., 1995) proteins. The increase in the proportion of integral membrane proteins in the larger genomes may suggest that communication with the outside world becomes relatively more important for cells in complex organisms.

The frequency distributions of proteins with different numbers of predicted transmembrane segments are shown in Figure 2. Although these distributions were calculated in three different ways, the general conclusions that the number of proteins falls off rapidly with increasing numbers of transmembrane segments and that transport proteins with 6 and 12 transmembrane segments are particularly prevalent in uni-cellular organisms whereas G-protein coupled receptors with seven transmembrane segments are characteristic of *C. elegans* and *H. sapiens*, do not depend on the details of the prediction method.

A notable result of the analysis is that integral membrane proteins of the helix-bundle class seem to come in two basic versions: those with a small number of transmembrane segments and large extra-membraneous domains, and those with many transmembrane segments and relatively small extra-membraneous domains (Fig. 3). Proteins with many transmembrane segments and large extra-membraneous domains are surprisingly rare. Indeed, in large membrane protein complexes such as cytochrome *c* oxidase, the extra-membraneous domains are in most cases made as separate, soluble subunits rather than as parts of the membrane-embedded subuntis (Iwata et al., 1995; Tsukihara et al., 1996).

The positive-inside rule is found to hold for eubacteria, eukaryotes, and the archaea *A. fulgidus* and *M. thermoautotrophicum*, while the results for the archaeon *M. jannaschii* are less clear (Fig. 4, Table 3). We, thus, cannot state for certain that the positive-inside rule can be extended to all organisms analyzed here: if it holds, it is in any case less extreme in *M. jannaschii*. Strictly speaking, the results presented in Figures 4 and 5 do not show that Arg and Lys residues are enriched in cytoplasmic segments of the proteins, only that there is a significant bias between the two sides of the membrane. For eubacterial and eukaryotic proteins with known topology, it has already been shown that it is the cytoplasmic segments that are enriched in positively charged residues. Regarding archaea, the only common protein where topology can be reliably inferred from sequence homology is SecY, and again, the cytoplasmic segments are strongly enriched in positively charged residues for all three species analyzed (Table 3). These observations imply that topology prediction methods based on the positive-inside rule (von Heijne, 1992; Jones et al., 1994; Persson & Argos, 1996; Rost et al., 1996) are applicable to membrane proteins from many if not all organism. Ideally, however, analyses such as the one presented in Figure 4 should be carried out on all newly sequenced genomes to assess the validity of the positive-inside rule before topology predictions are made.

Although negatively charged residues show only very slight biases, there is nevertheless a consistent trend in all organisms except *C. acetobutylicum* and *S. cerevisiae* that the net charge (Arg + Lys-Asp-Glu) has a slightly higher bias than the total charge (Arg + Lys + Asp + Glu), i.e., that positively and negatively charged residues tend to distribute to opposite sides of the topology (Fig. 5, top panel); this trend is particularly strong in *C. elegans* and *H. sapiens*. There is, thus, an indication of a weak "negative-outside" distribution, in addition to the strong "positive-inside" rule. Indeed, negatively charged residues have been shown to be able to affect membrane protein topology in both *E. coli* and mammalian systems (Spiess, 1995; Kiefer et al., 1997; Wahlberg & Spiess, 1997), although not as strongly as positively charged ones (Nilsson & von Heijne, 1990; Gafvelin et al., 1997).

In conclusion, a strong "positive-inside" and a weaker "negative-outside" distribution of residues is found in integral membrane proteins from all organisms studied so far, with the possible exception of *M. jannaschii*, where the results are not yet totally convincing. No other kind of residue shows a consistent bias in our analysis, although we cannot of course rule out weaker biases that are below the detection limit of the method used here. These observations suggest that the basic mechanisms of protein insertion into the membrane are quite similar in all living organisms, as attested to by the presence of the SecY/Sec61α translocon component—which is known to be involved in membrane protein assembly in *E. coli*, *S. cerevisiae*, and mammalian cells (Rapoport et al., 1996; de Gier et al., 1997)—in all so far sequenced genomes.

Methods

Amino acid bias analysis

All ORFs extracted from the genome sequences of *E. coli* (Blattner et al., 1997), *H. influenzae* (Fleischmann et al., 1995), *H. pylori* (Tomb et al., 1997), *Synechocystis* sp. (Kaneko et al., 1996), *B. subtilis* (Kunst et al., 1997), *C. acetobutylicum* (http://www.genomecorp.com/htdocs/sequences/clostridium/clospage.html), *M. genitalium* (Fraser et al., 1995), *M. pneumoniae* (Himmelreich et al., 1996), *M. jannaschii* (Bult et al., 1996), *M. thermoautotrophicum* (http://www.genomecorp.com/htdocs/sequences/methanobacter/abstract.html), *A. fulgidus* (Klenk et al., 1997), *S. cerevisiae* (Goffeau et al., 1997), *C. elegans* (sequences from SwissProt and TREMBL), and *H. sapiens* (sequences from SwissProt and TREMBL) were downloaded from the appropriate WWW site.

The TOPPRED program (Claros & von Heijne, 1994) was used to select sequences with at least seven "certain" candidate transmembrane segments (peak hydrophobicity above 1.0 using default parameters and the Engelman-Steitz hydrophobicity scale (Engelman et al., 1986)) and no "putative" candidate segments (peak hydrophobicity between 0.6 and 1.0). To avoid biasing the calculations by inclusion of long extra-membraneous segments we also required that all loops and the N- and C-terminal tails be $\leq$60 residues long.

Statistical analysis

To assess the degree of inside/outside amino acid bias in the selected proteins, we calculated for each kind of amino acid i and each protein k the absolute difference between the frequencies of the residue in the even (e)- and odd-numbered (o) extra-membraneous segments (see Fig. 4A), $\Delta f_{i,k} = |f^e_{i,k} - f^o_{i,k}|$. For each protein and each kind of amino acid, the corresponding absolute frequency difference was similarly calculated for all possible permutations of the extra-membraneous segments between the two sides of the structure, and the average absolute frequency difference $\langle \Delta f_{i,k} \rangle$ for the set of randomly permuted structures was determined. Finally, the difference between the observed bias and the average bias in the permuted set $\Delta F_{i,k} = \Delta f_{i,k} - \langle \Delta f_{i,k} \rangle$ was obtained, and these values were then averaged over all proteins from the organism in question and plotted (Fig. 4B).

To assess the statistical significance of the observed biases, we first calculated, for each kind of amino acid i and each protein k, the median $M_{i,k}$ of the absolute frequency differences in the permuted set and then recorded whether the observed absolute frequency difference $\Delta f_{i,k}$ for that protein was larger or smaller than the median. For each kind of amino acid i in each organism, we then estimated the probability P_i that the observed (n_i) or a higher number of proteins would have a bias larger than their corresponding $M_{i,k}$ from a binomial distribution as

$$P_i = \sum_{j=n_i}^{N} Bin(p = 0.5, N, j)$$

where *Bin* is the binomial distribution and N is the total number of proteins for that organism.

Prediction of transmembrane segments in the Sec61α/SecY protein family

Trans-membrane segments in the Sec61α/SecY family were predicted by a multiple-sequence extension of the TOPPRED method that in preliminary studies have performed significantly better than the original single-sequence algorithm (Wallin & von Heijne, in prep.). First, a multiple alignment was constructed using CLUSTAL W (Thompson et al., 1994), from which a consensus hydrophobicity profile was derived. Using the default parameters in TOPPRED (Claros & von Heijne, 1994), a unique topology with 10 transmembrane helices was predicted from the consensus hydrophobicity profile and was imposed on each of the individual family members in the alignment. The charge bias was then calculated for the individual proteins based on the consensus topology.

Acknowledgments

We are grateful to Dr. Douglas Smith and co-workers at Genome Therapeutics Corporation for providing unpublished sequence data for *C. acetobutylicum* and *M. thermoautotrophicum*. An anonymous reviewer suggested to study the relation between protein length and number of transmembrane segments (Fig. 3). This work was supported by grants from the Swedish Technical Sciences Research Council, the Swedish Natural Sciences Research Council, and the Göran Gustafsson Foundation to G.vH.

References

Akiyama Y, Ito K. 1987. Topology analysis of the SecY protein, an integral membrane protein involved in protein export in *Escherichia coli*. *EMBO J* 6:3465–3470.

Arkin IT, Brunger AT, Engelman DM. 1997. Are there dominant membrane protein families with a given number of helices? *Protein Struct Funct Genet* 28:465–466.

Beckmann R, Bubeck D, Grassucci R, Penczek P, Verschoor A, Blobel G, Frank J. 1997. Alignment of conduits for the nascent polypeptide chain in the ribsome–Sec61 complex. *Science* 278:2123–2126.

Blattner FR, Plunkett GR, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B, Shao Y. 1997. The complete genome sequence of *Escherichia coli* K-12. *Science* 277:1453–1462.

Bult CJ, White O, Olsen GJ, Zhou L, Fleischmann RD, Sutton GG, Blake JA, FitzGerald LM, Clayton RA, Gocayne JD, Kerlavage AR, Dougherty BA, Tomb JF, Adams MD, Reich CI, Overbeek R, Kirkness EF, Weinstock KG, Merrick JM, Glodek A, Scott JL, Geoghagen NSM, Venter JC. 1996. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* 273:1058–1073.

Claros MG, von Heijne G. 1994. TopPred II: An improved software for membrane protein structure prediction. *CABIOS* 10:685–686.

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

Cserzö M, Wallin E, Simon I, von Heijne G, Elofsson A. 1997. Prediction of transmembrane α-helices in prokaryotic membrane proteins: The Dense Alignment Surface method. *Protein Eng* 10:673–676.

de Gier J-WL, Valent Q, von Heijne G, Luirink J. 1997. The *E. coli* SRP: Preferences of a targeting factor. *FEBS Lett* 408:1–4.

Denzer AJ, Nabholz CE, Spiess M. 1995. Transmembrane orientation of signal-anchor proteins is affected by the folding state but not the size of the N-terminal domain. *EMBO J* 14:6311–6317.

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.

Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ, Kerlavage AR, Sutton G, Kelley JM, Fritchman JL, Weidman JF, Small KV, Sandusky M, Fuhrmann J, Nguyen D, Utterback TR, Saudek DM, Phillips CA, Merrick JM, Tomb J-F, Dougherty BA, Bott KF, Hu P-C, Lucier TS, Peterson SN, Smith HO, Hutchison CA, Venter JC. 1995. The minimal gene complement of *Mycoplasma genitalium*. *Science* 270:397–403.

Garavelin G, Sakaguchi M, Andersson H, von Heijne G. 1997. Topological rules for membrane protein assembly in eukaryotic cells. *J Biol Chem* 272:6119–6127.

Gavel Y, Steppuhn J, Herrmann R, von Heijne G. 1991. The positive-inside rule applies to thylakoid membrane proteins. *FEBS Lett* 282:41–46.

Gavel Y, von Heijne G. 1992. The distribution of charged amino acids in mitochondrial inner membrane proteins suggests different modes of membrane integration for nuclearly and mitochondrially encoded proteins. *Eur J Biochem* 205:1207–1215.

Gerstein M. 1997. A structural census of genomes: Comparing bacterial, eukaryotic, and archaeal genomes in terms of protein structure. *J Mol Biol* 274:562–576.

Goffeau A, Aert R, et al. 1997. The yeast genome directory. *Nature (Suppl)* 387:1–105.

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

Himmelreich R, Hilbert H, Plagens H, Pirkl E, Li BC, Herrmann R. 1996. Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acids Res* 24:4420–4449.

Hobohom U, Scharf M, Schneider R, Sander C. 1992. Selection of representative protein data sets. *Protein Sci* 1:409–417.

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.

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.

Kaneko T, Sato S, Kotani H, Tanaka A, Asamizu E, Nakamura Y, Miyajima N, Hirosawa M, Sugiura M, Sasamoto S, Kimura T, Hosouchi T, Matsuno A, Muraki A, Nakazaki N, Naruo K, Okumura S, Shimpo S, Takeuchi C, Wada T, Watanabe A, Yamada M, Yasuda M, Tabata S. 1996. Sequence analysis of

the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res 3*:109–136.

Kiefer D, Hu XT, Dalbey R, Kuhn A. 1997. Negatively charged amino acid residues play an active role in orienting the Sec-independent Pf3 coat protein in the *Escherichia coli* inner membrane. *EMBO J 16*:2197–2204.

Klenk H, Clayton R, Tomb J, White O, Nelson K, Ketchum K, Dodson R, Gwinn M, Hickey E, Peterson J, Richardson D, Kerlavage A, Graham D, Kyrpides N, Fleischmann R, Quackenbush J, Lee N, Sutton G, Gill S, Kirkness E, Dougherty B, McKenny K, Adams M, Loftus B, Peterson S, Reich C, McNeil L, Badger J, Glodek A, Zhou L, Overbeek R, Gocayne J, Weidman J, McDonald L, Utterback T, Cotton M, Spriggs T, Artiach P, Kaine B, Sykes S, Sadow P, D'Andrea K, Bowman C, Fujii C, Garland S, Mason T, Olsen G, Fraser C, Smith H, Woese C, Venter J. 1997. The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. *Nature 390*:364–370.

Kunst F, Ogasawara N, et al. 1997. The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. *Nature 390*:249–256.

Kutay U, Ahnerthilger G, Hartmann E, Wiedenmann B, Rapoport TA. 1995. Transport route for synaptobrevin via a novel pathway of insertion into the endoplasmic reticulum membrane. *EMBO J 14*:217–223.

Nielsen H, Engelbrecht J, Brunak S, von Heijne G. 1997. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng 10*:1–6.

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.

Persson B, Argos P. 1996. Topology prediction of membrane proteins. *Protein Sci 5*:363–371.

Persson B, Argos P. 1997. Prediction of membrane protein topology utilizing multiple sequence alignments. *J Protein Chem 16*:453–457.

Rapoport TA, Jungnickel B, Kutay U. 1996. Protein transport across the eukaryotic endoplasmic reticulum and bacterial inner membranes. *Annu Rev Biochem 65*:271–303.

Rost B, Fariselli P, Casadio R. 1996. Topology prediction for helical transmembrane proteins at 86% accuracy. *Protein Sci 5*:1704–1718.

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

Spiess M. 1995. Heads or tails—What determines the orientation of proteins in the membrane. *FEBS Lett 369*:76–79.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res 22*:4673–4680.

Tomb JF, White O, Kerlavage AR, Clayton RA, Sutton GG, Fleischmann RD, Ketchum KA, Klenk HP, Gill S, Dougherty BA, Nelson K, Quackenbush J, Zhou L, Kirkness EF, Peterson S, Loftus B, Richardson D, Dodson R, Khalak HG, Glodek A, McKenney K, Fitzegerald LM, Lee N, Adams MD, Hickey EK, Berg DE, Gocayne JD, Utterback TR, Peterson JD, Kelley JM, Cotton MD, Weidman JM, Fuji C, Bowman C, Watthey L, Wallin E, Hayes WS, Borodovsky M, Karp PD, Smith HO, Fraser CM, Venter JC. 1997. The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature 388*:539–547.

Tsukihara T, Aoyama H, Yamashita E, Tomizaki T, Yamaguchi H, Shinzawa-Itoh K, Nakashima R, Yaono R, Yoshikawa S. 1996. The whole structure of the 13-subunit oxidized cytochrome *c* oxidase at 2.8 Å. *Science 272*:1136–1144.

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

von Heijne G. 1992. Membrane protein structure prediction—Hydrophobicity analysis and the positive-inside rule. *J Mol Biol 225*:487–494.

von Heijne G. 1997. Principles of membrane protein assembly and structure. *Prog Biophys Mol Biol 66*:113–139.

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

Wahlberg JM, Spiess M. 1997. Multiple determinants direct the orientation of signal-anchor proteins: The topogenic role of the hydrophobic signal domain. *J Cell Biol 137*:555–562.

Wallin E, von Heijne G. 1995. Properties of N-terminal tails in G-protein coupled receptors—A statistical study. *Protein Eng 8*:693–698.

Wilkinson BM, Critchley AJ, Stirling CJ. 1996. Determination of the transmembrane topology of yeast Sec61p, an essential component of the endoplasmic reticulum translocation complex. *J Biol Chem 271*:25590–25597.