Comparative Analysis of Amino Acid Distributions in Integral Membrane Proteins From 107 Genomes

Johan Nilsson,[1,2,3] **Bengt Persson,**[1,2,3] **and Gunnar von Heijne**[3,4]*
[1]*Center for Genomics and Bioinformatics, Karolinska Institutet, Stockholm, Sweden*
[2]*IFM Bioinformatics, Linköping University, Linköping, Sweden*
[3]*Stockholm Bioinformatics Center, AlbaNova, Stockholm, Sweden*
[4]*Department of Biochemistry and Biophysics, Stockholm University, Stockholm, Sweden*

ABSTRACT **We have performed a comparative analysis of amino acid distributions in predicted integral membrane proteins from a total of 107 genomes. A procedure for identification of membrane spanning helices was optimized on a homology-reduced data set of 170 multi-spanning membrane proteins with experimentally determined topologies. The optimized method was then used for extraction of highly reliable partial topologies from all predicted membrane proteins in each genome, and the average biases in amino acid distributions between loops on opposite sides of the membrane were calculated. The results strongly support the notion that a biased distribution of Lys and Arg residues between cytoplasmic and extra-cytoplasmic segments (the positive-inside rule) is present in most if not all organisms. Proteins 2005;60:606–616.**

Key words: membrane protein; transmembrane helices; prediction; bioinformatics; hydrophobicity scale; genome-wide analysis

INTRODUCTION

Membrane proteins play crucial roles in all organisms, where they serve as, for example, receptors, ion channels, and various kinds of transport proteins. The vast majority of membrane proteins are formed by bundles of apolar transmembrane α-helices (TMHs). The fraction of open reading frames (ORFs) encoding α-helical membrane proteins in a typical genome has been estimated to 20–30%.[1–5] In contrast, less than 1% of the proteins of known 3D structure are membrane proteins.[6]

Given this huge gap between the number of known sequences and solved structures, computational methods for prediction of membrane protein structure play a crucial role for, for example, functional annotation and design of mutants. Although it is hard to predict their tertiary structure, the rather simple structural characteristics of membrane proteins allow for a relatively accurate prediction of their topology, that is, the number and positions of the apolar TMHs and the intracellular/extracellular locations of the polar connecting loops.

As far as is known, the strongest topological determinant in α-helical membrane proteins is the asymmetric distribution of positively charged residues between polar loops on opposite sides of the membrane.[7] This so-called positive-inside rule states that Lys and Arg residues are more prevalent in the nontranslocated than in the translocated parts of integral membrane proteins.

The positive-inside rule was first established for bacterial inner membrane proteins, mainly from *Escherichia coli*.[8] Subsequent studies on experimentally determined topologies confirmed that the rule applies also to eukaryotic plasma membrane proteins,[9] thylakoid membrane proteins,[10] and mitochondrial inner membrane proteins encoded within the mitochondrial genome.[11] However, the validity of the positive-inside rule has not so far been tested across a comprehensive set of organisms.

Here, we have sought to establish if the positive-inside rule is indeed universal, as is usually assumed. Predicted helical membrane proteins in 107 fully or partially sequenced genomes have been analyzed in terms of amino acid bias between the loops on opposite sides of the membrane. In summary, we find strong support for the notion that the positive-inside rule holds for most if not all organisms.

MATERIALS AND METHODS

Genome Databases

The predicted ORFs from a total of 107 genomes (15 archaea, 14 eukarya, and 78 eubacteria) were analyzed (see Electronic Supplement S2). For eubacterial and archaeal organisms, only completely sequenced genomes were included. For eukaryotes, the genomes of *Drosophila melanogaster*, *Caenorhabditis elegans*, *Arabidopsis thaliana*, *Plasmodium falciparum*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, and *Encephalitozoon cuniculi* were fully sequenced, while the remaining ones were only partially sequenced.

All eubacterial and archaeal ORFs were downloaded from ftp://ftp.ncbi.nlm.nih.gov. The eukaryotic ORFs of *Anopheles gambiae*, *Danio rerio*, *Caenorhabditis briggsae*,

---

The Supplementary Materials referred to in this article can be found at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/

Abbreviations: PT, partial topography; TMH, transmembrane helix; ORF, open reading frame.

*Correspondence to: Gunnar von Heijne, Department of Biochemistry and Biophysics, Stockholm University, SE-106 91 Stockholm, Sweden. E-mail: gunnar@dbb.su.se

Received 11 January 2005; Accepted 8 March 2005

Published online 18 July 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20583

and *Fugu rubripes* were downloaded from ftp://ftp.ensembl.org and the remaining eukaryotic ORF sets were downloaded from ftp://ftp.ncbi.nlm.nih.gov.

TMH Prediction

TMHs in the genome ORFs were identified using the hydrophobicity analysis (but not the charge-bias analysis) part of TopPred.[12,13] The TopPred algorithm constructs a hydrophobicity profile by averaging over a trapezoid sliding window using a given amino acid scale. Based on the resulting hydrophobicity profile, the algorithm then applies one upper and one lower cut-off value for TMH identification. Sequence segments corresponding to peaks in the profile above the upper cut-off are defined as "certain" TMHs, whereas peaks between the upper and lower cut-offs are designated as "putative" TMHs. In a final step, the algorithm then creates all possible topology models that can be obtained by always including all "certain" TMHs and either including or excluding each of the "putative" TMHs, and the positive-inside rule is applied to rank the different models.

Since we were only interested in the identification of protein segments containing TMHs that are predicted with a high reliability (and not full topologies), the final charge-bias ranking step in the TopPred algorithm was not applied. Instead, three parameters, the core window size, the wedge window size, and the hydrophobicity scale, were varied in an optimization procedure (see below) to obtain parameter combinations that identify TMHs with very high specificity in a set of proteins with experimentally determined topologies. The remaining user-adjustable TopPred parameters were kept at their default values. In the subsequent genome-wide analysis, the optimized parameter settings were used.

The following scales were evaluated in the optimization procedure: BT: Ben-Tal hydrophobicity scale,[14] EC: Eisenberg normalized consensus hydrophobicity scale,[15] GES: Goldman Engelman Steitz hydrophobicity scale,[16] HW: Hopp-Woods hydrophilicity value,[17] KD: Kyte-Doolittle hydropathy index,[18] WW: Wimley-White hydrophobicity scale,[19] RA: Rao & Argos membrane buried helix parameter scale,[20] and ZZ: Zhou & Zhou burial propensity scale.[21]

Partial topographies (PTs) were defined as regions in the predicted global topography consisting of at least five consecutive "certain" TMHs and no "putative" TMH. Hence, a single membrane protein may include more than one PT. The PT was defined to start at the N-terminal residue of the first certain TMH and end at the C-terminal residue of the last certain TMH. If the most N-terminal or most C-terminal TMH in a protein was part of a PT, the N-tail or C-tail of the protein was also included in the PT. Note that no in/out location is assigned to the loops in a PT.

Data Set of Membrane Proteins With Experimentally Determined Topologies

A data set of integral membrane proteins with experimentally characterized topologies was used both for amino acid bias analysis, and as a "positive training set" in the TMH prediction optimization procedure. The data was compiled from the Möller,[22] MPTOPO,[23] and TMPDB[24] databases, and from a number of recent publications. Furthermore, topologies were inferred for 13 recently determined X-ray structures of integral membrane proteins in the PDB.[25] For these proteins, the DSSP program[26] was used to determine the secondary structure of the proteins, and topologies were assigned by visual inspection using Rasmol.[27]

Only membrane proteins with two or more TMHs were included in the data set, and the set was subsequently homology reduced using an implementation of the Hobohm algorithm 2[28] with a pairwise global sequence identity threshold of 30%. Sequence alignments were generated using ClustalW.[29] If a protein was annotated to contain a cleavable signal peptide, in the data set used for TMH prediction optimization the corresponding region was removed, whereas in the data set used for amino acid bias analysis, the entire protein was discarded.

The data set used in the optimization process consisted of 170 proteins from all three domains of life, with 1203 TMHs in total. The data set is available as an Electronic Supplement (S1).

Data Set of Globular Proteins

A list of PDB identifiers representing a sequence-unique subset of 616 globular protein chains from the EVA server[30] was obtained from http://cubic.bioc.columbia.edu/papers/2002_htm_eval/data/. The protein sequences were obtained from the corresponding PDB entries, and were used as the "negative training set" in the optimization procedure.

TMH Prediction Optimization

Due to the limited size of the positive training set, a five-fold cross-validation was performed. The data set was divided into five equally sized subsets, of which four were used for training and the remaining one for testing. The average accuracy and coverage of the five cross-validation cycles was then calculated and used to evaluate each parameter combination. The accuracy of partial topography predictions was defined as the number of correctly predicted PTs divided by the total number of predicted PTs. The coverage was defined as the sum of TMHs in all predicted PTs divided by the sum of TMHs in all proteins in the test set. The aim was to determine, for each specific combination of parameter settings (hydrophobicity scale, and shape and size of the sliding window), the lowest upper cut-off value producing no false positive TMH predictions and the highest lower cut-off value producing no false negative TMH predictions. The optimization procedure was as follows:

1. *Determination of lower cut-off value*. The TMH predictor was used for predictions on the positive training set. The upper and lower cut-offs were initially set to values sufficiently low to assure that all TMHs in the data set were identified. A predicted TMH was considered a true positive if it overlapped with a TMH in the experimentally determined topology by at least nine residues. All

true positive TMHs were ranked according to the hydrophobicity score. The lowest-scoring segment was used to set the lower cut-off value, thus giving a minimal false negative prediction rate. TMHs with a score below three standard deviations from the mean were removed to avoid the influence of extreme values.

2. *Determination of upper cut-off value*. In a similar manner, TMH predictions were performed on the negative training set of globular proteins. The highest score obtained by a false positive TMH was taken as the upper cut-off value, allowing a minimal false positive prediction rate. Scores above three standard deviations from the mean of all false positive TMH scores were removed. The upper cut-off value was not subjected to cross-validation, and was thus kept constant in all five cycles.
3. *PT prediction evaluation*. The TMH predictor was rerun on the positive training set using the upper and lower cut-off values determined as described above. Partial topographies were extracted from the predictions and the accuracy and coverage was evaluated. The accuracy was obtained by comparing each predicted PT with the corresponding region in the experimentally determined global topology. The PT was considered correct if the number of TMHs matched and if each of the predicted TMHs overlapped with the corresponding helix in the experimental topology by at least three residues. The coverage was calculated as the fraction of all TMHs in the partial topographies and all TMHs in the entire data set.
4. *Average performance estimation*. To obtain a final estimate of the performance of a certain parameter combination, the average prediction accuracy and coverage from the five cross-validation cycles was calculated. The final parameter combinations were chosen as presented under Results.

Genome Predictions

TMH identification using the optimized parameter settings was performed on all predicted membrane proteins in each genome. PTs were extracted from the identified TMHs as described above and, for each genome, the full set of amino acid sequences representing the PTs was subjected to homology reduction using the Hobohm algorithm 2 with a sequence identity threshold of 30%. The resulting homology-reduced partial topographies were then checked for biases in the distribution of each of the 20 amino acids between loops on opposite sides of the membrane, as described previously.[1] Briefly, for each amino acid residue type, i, and each PT, k, the absolute frequency difference, $\Delta f_{i,k}$, between the odd- and even-numbered loops was calculated. Subsequently, all possible permutations of the loop segments in a given PT were generated and the corresponding absolute frequency differences, $\Delta f'_{i,k}$, were calculated for each permutation. The mean of the frequency differences for the permuted PTs, $\langle \Delta f'_{i,k} \rangle$, was then calculated and subtracted from the observed absolute frequency difference, $\Delta f_{i,k}$, in the nonpermuted PT. These values $(\Delta f_{i,k} - \langle \Delta f'_{i,k} \rangle)$ were finally averaged for each amino acid residue type over all PTs from the organism in question, to yield an estimation of the bias.

To assess the statistical significance of an observed bias, the median value, $M_{i,k}$, of the absolute frequency differences, $\Delta f'_{i,k}$, in all permutations was first calculated for each PT. For each amino acid residue type, i, the number of PTs, n_i, where the observed absolute frequency difference, $\Delta f_{i,k}$, was larger than the median absolute frequency difference in the permuted set, $M_{i,k}$, was then counted. The probability, P_i, to observe n_i or more PTs with a bias higher than $M_{i,k}$, given the hypothesis that i is equally frequent on either side of the membrane, was assessed from a binomial distribution. An observed bias was considered significant if the P_i value was below 1%.

Analysis of Proteins With Known Topology

The data set of membrane proteins with experimentally determined topologies described above was divided into one prokaryotic and one eukaryotic subset. The two data sets were then homology reduced separately using the Hobohm algorithm 2, resulting in a homology-reduced prokaryotic data set of 114 proteins and a homology-reduced eukaryotic data set of 63 proteins. The frequency bias of each amino acid residue type between cytoplasmic and extra-cytoplasmic polar segments was calculated and averaged over all eukaryotic and prokaryotic proteins, respectively. The significance of an observed bias in amino acid distribution between cytoplasmic and extra-cytoplasmic loops was assessed in two ways: by χ^2 analysis, and by a simple binomial test. In the χ^2 test, the frequency of each amino acid residue type, i, was investigated in cytoplasmic and extra-cytoplasmic loops, respectively, under the null hypothesis that i is equally frequent in loops on either side of the membrane. In the binomial test, the number of proteins, n_i, where a certain amino acid residue type, i, was more frequent in cytoplasmic than extra-cytoplasmic segments (or vice versa) was counted. The probability to observe such a bias in n_i or more proteins was then calculated from the binomial distribution, given the same null hypothesis as above. The binomial test, while being less sensitive than χ^2 analysis, avoids the influence of proteins where a certain bias might be very pronounced.

The preference of Lys and Arg residues to cluster near the ends of long loop regions was investigated in all inter-TMH regions of at least 60 residues length in the prokaryotic and eukaryotic data sets. The number of cytoplasmic loops of required length was 18 in the prokaryotic and 28 in the eukaryotic set, and the corresponding number of extra-cytoplasmic loops was 22 and 18. For each of these four loop sets, the frequency of (Lys+Arg) residues was calculated for the 10 most N-terminal, the 10 most C-terminal and the 20 central residues. The compositional biases between the different parts were then assessed by χ^2 analysis. In addition, the mean frequency of positively (Lys+Arg) and negatively (Asp+Glu) charged residues in cytoplasmic and extra-cytoplasmic loops flanking the first, second, etc., TMH was calculated and averaged over all prokaryotic and eukaryotic proteins, respectively. Polar

segments longer than 60 residues were excluded from the analysis.

The net charge difference ($n_{Arg} + n_{Lys} - n_{Asp} - n_{Glu}$) between the 10–15 residues flanking the most N-terminal TMH was also calculated for each eukaryotic protein. In a given protein, if either of the flanking loops was shorter than 10 residues, the loop pair was excluded from the analysis. If both loops were 15 residues or longer, the 15 flanking residues were analyzed and if one loop was 15 residues or longer and the other was k residues, where $10 < k < 15$, the k flanking residues were analyzed. Using these criteria, the net charge bias was analyzed in a total of 40 homology-reduced eukaryotic proteins.

Analysis of the SecY/Sec61α Protein Family

The three dimensional structure of SecY from *Methanococcus jannaschii* is known at atomic resolution.[31] SecY orthologs in the other genomes were identified by either searching for annotated SecY/Sec61α sequences in SwissProt or, if no SwissProt entry was found, by performing a FASTA search of the *M. jannaschii* and *E. coli* SecY sequences against all predicted ORFs from the genome. Multiple sequence alignments were built from the sequence of the *M. jannaschii* SecY protein, and the orthologs from each domain of life (archaea, eubacteria, and eukarya) using ClustalW with default settings. The topology of each ortholog was then inferred by imposing the known topology of *M. jannaschii* SecY on the multiple sequence alignment. The frequency difference of each amino acid residue type between cytoplasmic and extracytoplasmic loop regions was calculated for each SecY ortholog.

RESULTS

Previous theoretical and experimental studies have shown that a biased distribution of certain amino acid residue types plays an important role in determining the topology of a membrane protein.[32] The most well-established bias is the enrichment of Lys and Arg residues in the nontranslocated polar parts of a protein, the so-called positive–inside rule. The aim of this study was to investigate if the positive–inside rule holds for membrane proteins in a number of organisms with fully or partially sequenced genomes.

The basic idea behind the analysis presented here is similar to that employed in a previous study.[1] The objective was to include in the amino acid bias analysis only those proteins in a given genome for which a simple hydrophobicity analysis, with very strict cut-offs for the identification of TMH and non-TMH segments, unambiguously predicts only one topography (by "topography" we mean a specification of the TMHs but without extracellular/intracellular orientational information). After this initial selection step, amino acid composition biases between the two sides of the predicted topography were calculated for all proteins and averaged over the whole genome. In contrast to the previous work, where only full unambiguous topographies were considered, we now include unambiguous partial topographies (PTs), defined as parts of proteins that contain at least five consecutive, unambiguously predicted TMHs. This allows us to considerably expand the number of proteins included in the analysis, and to perform homology-reduction on the data set prior to the compositional bias calculation. It is important to note that the positive–inside rule or other indices of amino acid bias between the intra- and extracellular parts of a membrane protein are included either explicitly or implicitly in nearly all currently used topology prediction algorithms, and we thus cannot employ any of these methods in our derivation of PTs. Instead, we use only a simple hydrophobicity analysis protocol to identify TMHs that are predicted with a very high reliability without including any information relating to intra/extracellular amino acid biases. These reliably predicted TMHs are then used to establish the partial topographies upon which the subsequent amino acid bias analysis is based.

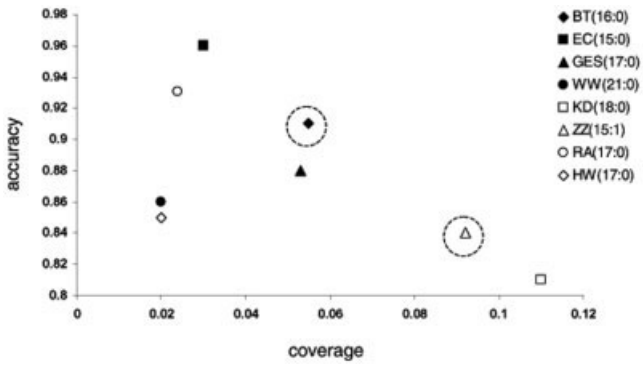

Fig. 1. Partial topography prediction coverage and accuracy for the eight hydrophobicity scales. Each hydrophobicity scale is designated with its abbreviation (see Materials and Methods) followed by the core window size and the wedge window size. The coverage was calculated as the quotient of the total number of TMHs in all predicted PTs and the total number of TMHs in all test set proteins. The accuracy is the fraction of correctly predicted partial topographies. The two parameter combinations that were chosen for the genome analysis are indicated with circles.

Hydrophobicity Analysis Optimization

As a first step, we sought to determine which parameter combinations to use in the hydrophobicity analysis by optimizing over a homology-reduced training set of 170 prokaryotic and eukaryotic proteins with experimentally determined topologies. The hydrophobicity analysis optimization was done with the TopPred algorithm using eight different hydrophobicity scales and a trapezoid sliding window described by two parameters: the size of the rectangular core window, and the sizes of the flanking wedge windows. The homology-reduced data set of 170 proteins with experimentally determined topologies is described in Materials and Methods; it contains 109 eubacterial, four archaeal, and 57 eukaryotic proteins. Fivefold cross-validation was used in the parameter optimization step.

Figure 1 shows the mean prediction accuracy and coverage for the best-performing parameter settings of a given hydrophobicity scale. The graph shows a roughly linear inverse relationship between prediction accuracy and test

set coverage for the best-performing scales. Based on these results, two parameter combinations were chosen for the genome analysis: the Ben-Tal hydrophobicity scale with a rectangular window of 16 residues and no wedge window (designated BT(16:0) in Fig. 1) and the Zhou-Zhou scale of mean burial propensity with a core window size of 15 and a wedge window size of 1 (designated ZZ(15:1) in Fig. 1). The upper and lower cut-off values were 0.45 and −0.97, respectively, for the Ben-Tal scale, and 0.79 and 0.75, respectively, for the Zhou-Zhou scale. While the BT(16:0) protocol yields a lower false-positive rate, the ZZ(15:1) protocol has almost twice the coverage. Thus, the two scales should complement each other. More specifically, for analysis of larger genomes BT(16:0) should be more informative than ZZ(15:1), since it is more accurate and still gives a considerable number of predicted PTs. For smaller genomes, ZZ(15:1) should be more informative thanks to its better coverage.

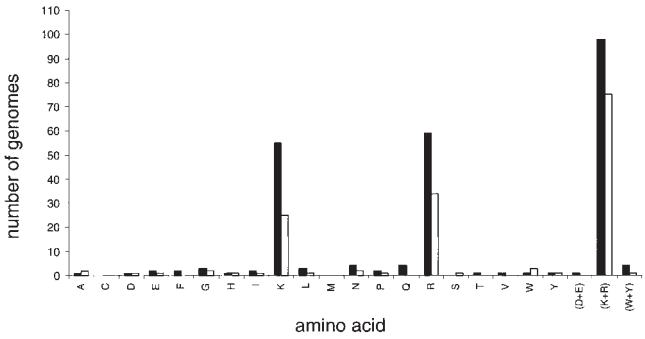

Fig. 2. Number of genomes where a certain amino acid residue type is biased according to the binomial test. White bars indicate predictions performed using the BT(16:0) protocol, and black bars indicate predictions using the ZZ(15:1) protocol. Amino acids are indicated by their single letter abbreviations. In addition to the individual amino acids, the bias for (Asp+Glu), (Lys+Arg), and (Trp+Tyr) are displayed.

Genome Predictions

All predicted ORFs in each genome were subjected to hydrophobicity analysis using the two scale/parameter setting combinations discussed above. Subsequent extraction of PTs and homology reduction was then performed as described in Materials and Methods. Unless otherwise specified, all results presented are for predictions performed using the ZZ(15:1) protocol.

When the eubacterial and archaeal PTs were subjected to homology reduction, in most cases only between 10–30% of the PTs were removed. In contrast, up to 50% of the eukaryotic PTs were removed in this step (data not shown). At least for the latter genomes, the homology reduction step is thus critically important in order to avoid bias from related proteins in the subsequent analysis.

All genomes analyzed, along with the total number of ORFs, the total number of integral membrane proteins predicted by TMHMM2.0 and the total number of PTs (after homology reduction) obtained using either the BT(16:0) or ZZ(15:1) protocol are listed in Electronic Supplement S2.

Amino Acid Biases in Genomes

The inside/outside bias for a given type of amino acid residue, i, in a given PT, k, was calculated as described in Wallin and von Heijne,[1] that is, by comparing the observed absolute frequency difference, $\Delta f_{i,k}$, with the average absolute frequency difference calculated for all possible permutations of loops for that PT, $\langle \Delta f'_{i,k} \rangle$. The difference $\Delta f_{i,k} - \langle \Delta f'_{i,k} \rangle$ was then averaged over all PTs in the genome and a simple statistical test was applied (see Materials and Methods for details). Finally, the amino acid residue types were ranked in order of decreasing average bias for each genome in order to discern common trends.

The (Lys+Arg) bias is significant on the 1% level for 98 of the 107 genomes using the ZZ scale and for 75 genomes using the BT scale (see Electronic Supplement S2). Taken individually, Lys is significantly biased in 55 genomes using the ZZ scale and 25 genomes using the BT scale, and the corresponding figures for Arg are 59 and 34. In contrast, the remaining amino acid residue types are significantly biased in only a few or no genomes (see Fig. 2). Besides, with the one exception of Tyr in *Clostridium acetobutylicum*, no other amino acid residue type than Lys or Arg is consistently biased with both hydrophobicity scales in a genome (see Electronic Supplement S2).

Given the rather small number of PTs in many genomes (especially when using the BT scale), it is not surprising that statistical significance is not always seen. For this reason we also performed a rank analysis to check if Arg and Lys are consistently the most biased amino acid residues even in genomes where the (Lys+Arg) bias is not significant on the 1% level. Figure 3 shows the number of genomes in which a certain amino acid residue type is ranked in a certain order. In all but five of the 107 genomes analyzed, the top-ranked bias is for (Lys+Arg), and individually one of these two amino acids also occupies the second rank in 99 of the genomes. However, there is a notable difference in the magnitude of the Lys and Arg bias between prokaryotes and eukaryotes (Fig. 4). Thus, while proteins in archaeal and eubacterial organisms display a similar (Lys+Arg) bias (the median bias is 4.2% for both eubacteria and archaea), this bias is clearly weaker in eukaryotes (median bias 2.0%); the same trend is seen when Arg and Lys biases are analyzed separately. The negatively charged residues Glu and Asp, on the other hand, clearly rank below Arg and Lys (Fig. 3) and their mean bias across all genomes is close to zero, see Figure 4.

To check the robustness of the results from the genome analysis, we varied the minimum number of loops required to define a PT between 4 and 7. Although this caused small fluctuations in the calculated biases, the major trends remained (i.e., Arg or Lys is top-ranked in almost all genomes, data not shown). We have used a minimum value of five loops in all results presented.

Amino Acid Biases in Known Topologies

To investigate if a similar trend was observed in experimentally determined topologies, we calculated the mean inside/outside bias for each amino acid residue type in two homology-reduced data sets consisting of 114 prokaryotic

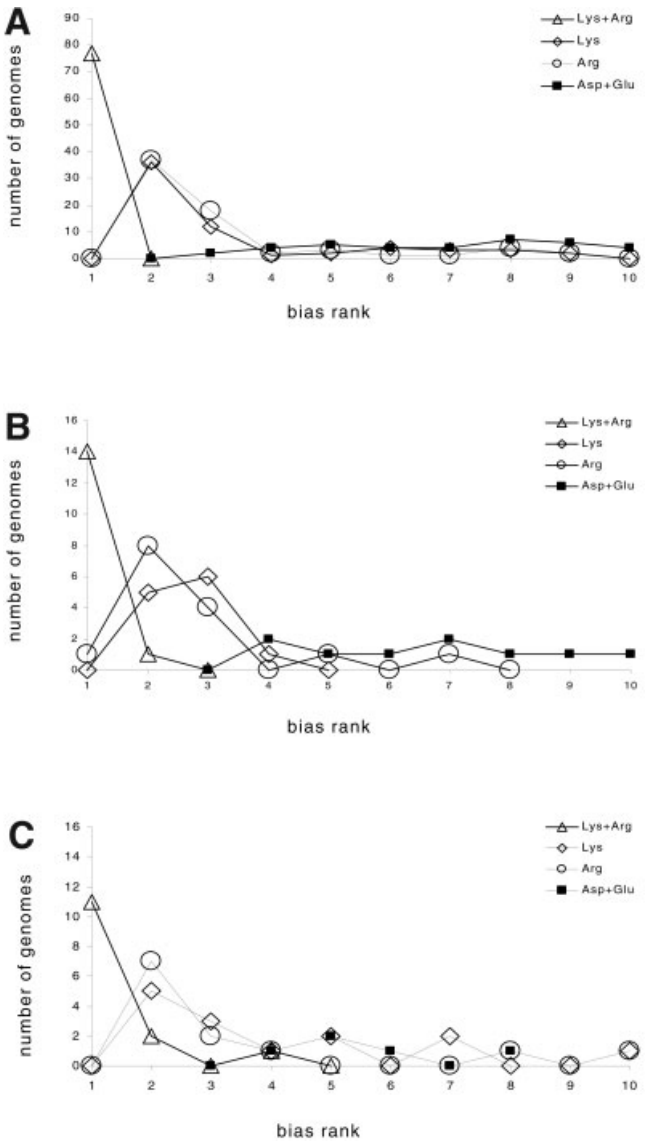

Fig. 3. Bias rank of a certain amino acid residue type in (**A**) eubacteria, (**B**) archaea and (**C**) eukaryotes. A bias rank of 1 means this amino acid has the highest frequency bias between intracellular and extracellular loops, a rank of 2 means it has the second highest bias etc. All PT predictions were performed using the ZZ(15:1) protocol.

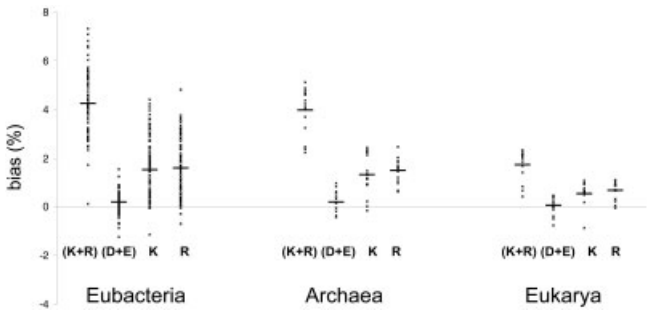

Fig. 4. Mean inside/outside bias of (Lys+Arg), (Asp+Glu), Lys and Arg in partial topographies from all eubacterial, archaeal, and eukaryotic organisms. Each cross corresponds to the bias of the indicated amino acid residue type in a certain genome belonging to one of the three domains of life. The mean bias of an amino acid residue type *i* in a genome is calculated as described in Materials and Methods. The mean bias for the entire domain is indicated with a horizontal bar. All partial topography predictions were performed using the ZZ(15:1) protocol.

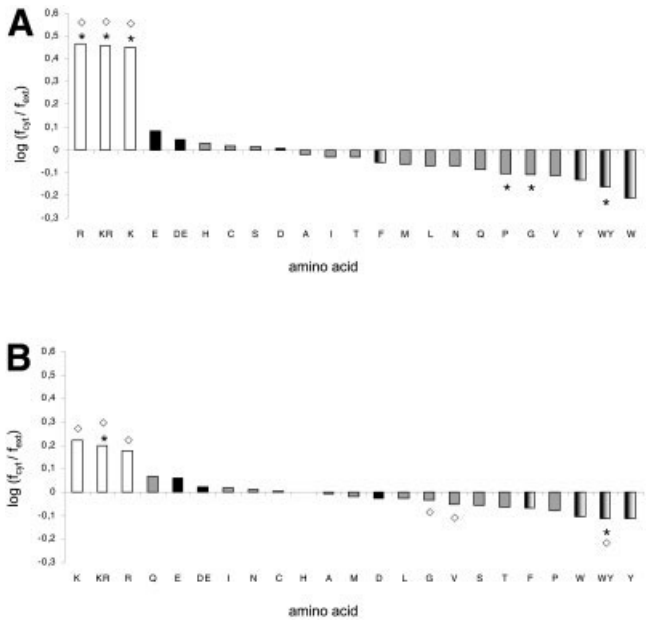

Fig. 5. Frequency bias in proteins with known topology. The bar chart shows the frequency bias of each amino acid residue type between cytoplasmic and extra-cytoplasmic polar segments in a sample of (**A**) 114 prokaryotic and (**B**) 63 eukaryotic multi-spanning membrane proteins. Amino acids are designated by their one-letter abbreviations (DE = Asp + Glu, KR = Lys + Arg and WY = Trp + Tyr). A positive value indicates a prevalence of the amino acid residue type for cytoplasmic segments and a negative value indicates a prevalence for extra-cytoplasmic segments. Biases that are significant ($P < 0.01$) as assessed by a binomial test are indicated with an asterisk at the respective bar, whereas biases significant by the χ^2 test ($P < 0.001$) are indicated with a diamond. White bars designate positively charged amino acids (Arg, Lys), black bars designate negatively charged amino acids (Asp, Glu), vertically shaded bars designate aromatic amino acids (Phe, Trp, Tyr), and grey bars designate the remaining amino acid residue types.

proteins and 63 eukaryotic proteins, respectively. Figure 5 shows the mean frequency bias of each amino acid residue type between cytoplasmic and extra-cytoplasmic loops. The significance of observed biases was assessed by a binomial test as described in Materials and Methods. It is obvious also here that there is a strong accumulation of Lys and Arg in cytoplasmic loops of prokaryotic proteins, while this bias is clearly weaker in the eukaryotes. Also, the frequency of aromatic residues (Trp+Tyr) is significantly higher in extra-cytoplasmic loops of both prokaryotic and eukaryotic proteins. In contrast, no "negative-outside" bias (i.e., accumulation of Asp and Glu in extra-cytoplasmic loops) is apparent.

To get an idea whether the observed (Lys+Arg) bias is always due to a prevalence of these amino acid residues in cytoplasmic loops, and not the opposite (i.e., a reverse, or "positive-outside," distribution), we searched the COG database[33] for membrane proteins present in most or all of the investigated organisms. The only such protein family we could find where the topology was experimentally determined was SecY/Sec61α; the X-ray structure of SecY

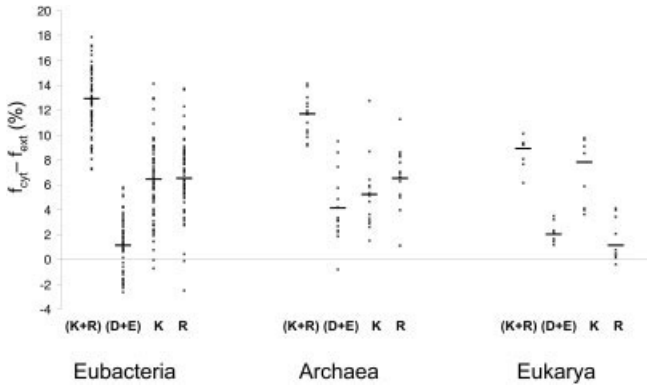

Fig. 6. Analogous to Figure 4, but results were obtained from analysis of the SecY/Sec61α protein in each organism rather than from predicted partial topographies. Also, since the orientation is known for this protein, the bias displayed is the frequency difference of the residue between the cytoplasmic and extra-cytoplasmic polar segments.

TABLE I. Amino Acid Bias in SecY/Sec61α†

Organism	$f_{cyt} - f_{ext}$ (%) (Arg + Lys)	Lys	Arg	(Asp + Glu)
Eubacteria				
BORBU	8.9	9.0	−0.1	0.9
BUC	14.3	11.0	3.3	−0.9
BUCAI	10.8	5.2	5.6	1.4
CHLCV	13.5	4.8	8.7	1.8
RICPR	16.8	12.9	3.9	0.7
Archaea				
METKA	14.1	2.9	11.2	8.6
Eukarya				
ENCCU	—	—	—	—
PLAFA	8.1	4.0	4.1	2.2
YEAST	6.2	4.1	2.1	3.5

†Frequency difference for charged residues between cytoplasmic and extra-cytoplasmic polar segments in the SecY/Sec61α protein from organisms that did not show a significant Arg or Lys bias in the genome analysis.
Abbreviations: BORBU, *Borrelia burgdorferi;* BUC, *Buchnera aphidicola, subsp. S. graminum;* BUCAI, *Buchnera aphidicola, subsp. A. pisum;* CHLCV, *Chlamydophila caviae;* RICPR, *Rickettsia prowazekii;* METKA, *Methanopyrus kandleri;* ENCCU, *Encephalitozoon cuniculi;* PLAFA, *Plasmodium falciparum;* YEAST, *Saccharomyces cerevisiae.*

from *Methanococcus jannaschii* was recently determined.[31] The topology of orthologs in the remaining genomes was inferred from the *M. jannaschii* SecY topology by multiple sequence alignments (see Materials and Methods). There were two eukaryotic genomes (*D. rerio* and *E. cuniculi*) for which the SecY/Sec61α topology could not be reliably inferred from the multiple sequence alignment.

Figure 6 shows the mean amino acid frequency difference between cytoplasmic and extra-cytoplasmic polar segments in all genomes where a SecY/Sec61α ortholog could be identified. The trend is basically the same as in the absolute frequency bias analysis of complete genomes, that is, there is a clear prevalence of positively charged residues in the cytoplasmic loops. Likewise, the bias magnitude is lower in eukaryotes than in archaea or eubacteria, although the difference is less pronounced compared to the complete genomes. The fact that all of the analyzed SecY/Sec61α orthologs display a higher (Lys+Arg) frequency in cytoplasmic loops suggests that none of the investigated organisms have a "positive-outside" distribution.

The frequency differences of positively and negatively charged residues between cytoplasmic and extra-cytoplasmic loops in the SecY/Sec61α protein from genomes where neither Lys nor Arg were significantly biased according to the binomial test are shown in Table I. The biases appear to be of roughly the same magnitude as the corresponding biases in the remaining genomes (cf. Fig. 6). Interestingly, in most of the orthologs the mean frequency of the negatively charged Asp and Glu residues is higher in cytoplasmic than extra-cytoplasmic segments (Fig. 6). The mean frequency difference is similar for Arg and Lys in eubacterial and archaeal SecY, whereas Arg displays a much weaker bias than Lys in Sec61α from eukaryotes (mean frequency difference of 1.1% vs. 7.8%). However, this might simply be an effect of the relatively low number of eukaryotic orthologs.

Previous studies have suggested that the net charge difference over the most N-terminal membrane-spanning segment in eukaryotic (but not prokaryotic) membrane proteins correlates with its orientation in the membrane (the so-called "charge difference" rule), and that this bias is the result of an asymmetric distribution of both positively and negatively charged residues.[9,34] To investigate the contribution of positively and negatively charged residues to the charge difference rule, we analyzed specifically the distribution bias of these residues in cytoplasmic and extra-cytoplasmic segments flanking the most N-terminal TMH in the homology-reduced data sets of prokaryotic and eukaryotic proteins with known topology. Figure 7 shows the mean frequency of (Asp+Glu) and (Lys+Arg) in cytoplasmic and extra-cytoplasmic segments as a function of the position of the segment, counting from the N-terminus. As noted previously,[9] the biased distribution of positively charged residues is roughly uniform throughout both prokaryotic and eukaryotic proteins [Fig. 7(A,C)]. Interestingly, the frequency difference of (Asp+Glu) over the most N-terminal TMH in the eukaryotic proteins (11.1% *vs.* 9.0%, Fig. 7d) is considerably less pronounced here than in the previous studies, possibly due to the fact that our data set is larger and has been homology reduced. Furthermore, when the charge bias over the most N-terminal TMH was analysed in a total of 40 eukaryotic proteins, the correct N-tail location could be predicted for 27 proteins by analysis of the net charge bias, whereas the positive-inside rule alone predicted the correct N-tail location in 28 proteins (data not shown). The "negative outside" bias in the N-terminal parts of eukaryotic membrane proteins reported earlier thus seems to be weak at best and probably has only a minor influence on the topology of most proteins.

The positive-inside rule is usually assumed to hold only for rather short loops (less than 60 residues). A tendency of positively charged residues to cluster near the ends of cytoplasmic loops has been noted in previous studies.[35,36]

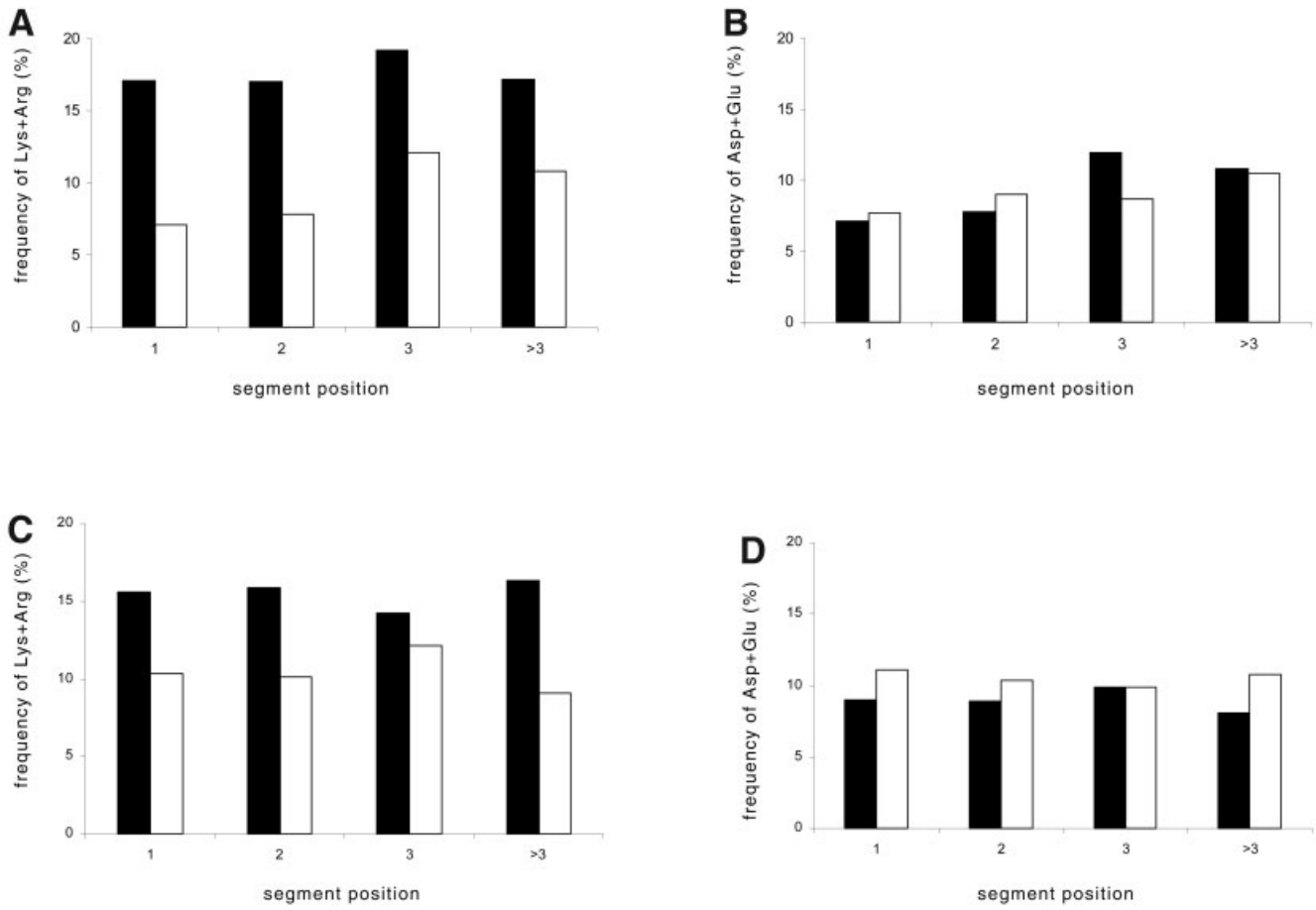

Fig. 7. Frequency of (Lys+Arg) and (Asp + Glu) in 114 prokaryotic (**A**, **B**) and 63 eukaryotic (**C**, **D**) membrane proteins of known topology as a function of the position of the segment, counting from the N-terminus. Black bars designate cytoplasmic polar segments and white bars designate extra-cytoplasmic polar segments.

However, to our knowledge, no study has specifically investigated the clustering in long loops. Figure 8 shows the frequency of (Lys+Arg) residues in the N-terminal, C-terminal, and central parts of loops of at least 60 residues length. It is clear from the figure that a strong accumulation of positively charged residues occurs near the ends of cytoplasmic loops in prokaryotic proteins. The frequency difference between the central part of the loops and either the N-terminal or C-terminal part of the loops is significant as assessed by χ^2 analysis ($P < 0.001$). However, there seems to be no corresponding bias in long loops from eukaryotic proteins [Fig. 8(B)].

DISCUSSION

The main topological determinants of integral membrane proteins are considered to be the long, apolar stretches that form the membrane spanning α-helices, and the biased distribution of Arg and Lys residues in polar regions exposed on either side of the membrane, the so-called positive-inside rule. In this study we have assessed the validity of the positive-inside rule by analyzing the amino acid composition bias in predicted membrane proteins from 107 organisms.

As the initial step, we used a simple hydrophobicity analysis protocol without reference to the positive-inside rule to identify transmembrane segments, for two reasons. First, since the main objective was to investigate which organisms follow the positive-inside rule, it was crucial that the predicted partial topographies (PTs)—specifications of transmembrane segments without orientational information—were not themselves based on this rule in any way. Second, since the analysis was performed on large-scale genomic data, we could allow for the loss of coverage resulting from the strict criteria used in the analysis, and also perform homology reduction before the final analysis.

The hydrophobicity criteria were chosen based on a procedure where we sought to find an optimal combination of hydrophobicity scale and window size validated against a large collection of membrane proteins with experimentally known topologies. From this analysis, we chose two different combinations: one that predicts PTs with very high accuracy but has a low coverage, and one with a somewhat lower accuracy but considerably better coverage.

By performing the analysis on partial rather than full topographies, the sensitivity is increased compared to a

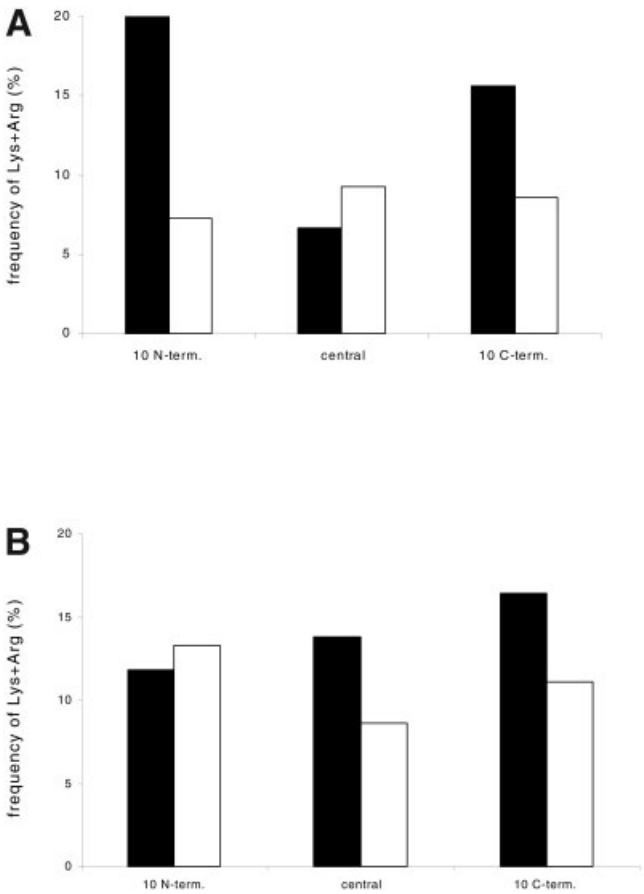

Fig. 8. Frequency of (Lys+Arg) in loops of at least 60 residues length from prokaryotic (**A**) and eukaryotic (**B**) membrane proteins of known topology. Frequencies were calculated separately for the 10 most N-terminal, the 10 most C-terminal and the 20 central residues in all loops. Black bars designate cytoplasmic loops and white bars designate extracytoplasmic loops.

previous study.[1] Hence, a significant (Lys+Arg) bias is detected even in such small eubacterial genomes as that of the *Mycoplasma* species (*M. penetrans* and *M. pulmonis*). Table II lists the top ranked bias in genomes for which neither of the two TMH predictions using the optimized parameter settings gave a significant Arg or Lys bias. The fact that Arg or Lys is the top ranked bias in all but two of these genomes suggests that the lack of statistical significance is due to the limited number of PTs.

The bias of Lys and Arg residues between intra- and extracellular loop regions is well established for eubacterial membrane proteins (mainly those of *E. coli*) as well as for eukaryotic proteins, where the positive-inside rule has been confirmed in several studies on experimental topologies. However, the lack of experimental data for archaeal membrane proteins has not allowed for a similar analysis of the charge bias in these organisms. In this study, analysis of the frequency bias of different amino acid residue types in predicted partial topographies from 15 archaeans reveals a bias pattern very similar to that of eubacteria.

For the 15 archaeans, a significant (Lys+Arg) bias is obtained with either or both of the amino acid scales in all genomes except that of *Methanopyrus kandleri* (see Electronic Supplement S2). This genome is one of the smallest (1687 ORFs in total) and it has a substantially lower number of predicted PTs compared to the other archaeal organisms. Moreover, Lys is the top ranked bias using the BT scale and (Lys+Arg) is top ranked when predictions are performed with the ZZ scale on this genome (Table II). Considering this, and the fact that the organism belongs to phylum *Euryarchaeota*, to which the majority of the other archaeans studied here also belong, it seems likely that the positive-inside rule holds also for this genome.

Overall, only six out of 94 prokaryotic genomes did not show a statistically significant (Lys+Arg) bias with any of the two scales on the 1% level by the binomial test we applied (see Table II). As expected, all of these genomes are relatively small, and with the exception of *Buchnera aphidicola*, subsp. *S. graminum*, the (Lys+Arg) bias is top-ranked in all genomes, using either the BT or ZZ scales (Table II). The enterobacterium *Buchnera aphidicola* is an endosymbiont of aphids and two strains have been fully sequenced: the symbiont of *Schizaphis graminum* and that of *Acyrthosiphon pisum*. The fact that the (Lys+Arg) bias is top ranked in the symbiont of *A. pisum*, suggests that the bias is likely to be present also in the second strain. Also, the (Lys+Arg) frequency bias is relatively high in SecY from the *S. graminum* symbiont (14.3%, see Table I).

The situation is slightly different for the eukaryotic genomes. Here, no significant Lys or Arg bias was detected in three of the 15 genomes (Table II). Also, it is evident from Figure 4 that the magnitude of the bias is lower in eukaryotes. This latter finding is in accordance with previous studies on experimental topologies of eukaryotic membrane proteins.[9,37] Nevertheless, the Lys or Arg bias is top-ranked in all eukaryotic genomes except that of *Plasmodium falciparum*, using either the BT or ZZ scales (Table II).

While earlier results[1] suggested that Lys tends to be more biased than Arg, our results show a surprisingly equal bias for these two residue types. For instance, when predictions were performed using the ZZ scale, in eubacteria Arg and Lys both displays a significant bias in 42 of the genomes. For predictions with the BT scale, the corresponding figures are 28 genomes for Arg and 20 genomes for Lys. Also, the mean and median bias magnitudes averaged over all eubacterial genomes are virtually equal (mean bias of 1.5% for Lys and 1.6% for Arg and median biases of 1.5% for Lys and 1.4% for Arg).

The topology of the strongly conserved SecY/Sec61α protein has been determined at atomic resolution. This allows for a more accurate analysis of the amino acid distributions, since the borders of the membrane spanning segments can be more precisely defined. In all genomes for which the SecY/Sec61α topology could be inferred from the known X-ray structure, a strong bias in the positively charged residues can be seen (Fig. 6). In agreement with the results obtained from analysis of predicted topographies, the Sec61α protein from eukaryotes has a lower

TABLE II. Bias Rank in Genomes Where (Lys+Arg) Bias is Not Significant[†]

Organism	Total nr of ORFs	Total nr of membrane proteins (≥2TMHs)	Total nr of PTs, ZZ scale	Total nr of PTs, BT scale	Top-ranked aa type, ZZ scale	Top-ranked aa type, BT scale
Eubacteria						
BORBU	1637	154	26	9	(K+R)	(K+R)
BUC	504	61	14	9	N	(W+Y)
BUCAI	574	58	15	4	(K+R)	F
CHLCV	1005	187	30	6	(K+R)	(K+R)
RICPR	834	148	34	15	(K+R)	(K+R)
Archaea						
METKA	1687	198	20	6	(K+R)	K
Eukarya						
ENCCU	1996	262	46	1	N	K
PLAFA	5267	826	71	14	L	L
YEAST	5855	718	93	12	S	(K+R)

[†]Top bias ranked amino acid residue type in genomes where neither of the two amino acid scales yielded a statistically significant Arg or Lys bias by the binomial test.
Abbreviations: BORBU, *Borrelia burgdorferi;* BUC, *Buchnera aphidicola, subsp. S. graminum;* BUCAI, *Buchnera aphidicola, subsp. A. pisum;* CHLCV, *Chlamydophila caviae;* RICPR, *Rickettsia prowazekii;* METKA, *Methanopyrus kandleri;* ENCCU, *Encephalitozoon cuniculi;* PLAFA, *Plasmodium falciparum;* YEAST, *Saccharomyces cerevisiae.*

(Lys+Arg) bias between cytoplasmic and extra-cytoplasmic segments than prokaryotic SecY. Nevertheless, the average (Lys+Arg) bias for Sec61α is 8.9%. The frequency difference of (Lys+Arg) between cytoplasmic and extracytoplasmic loops in Sec61α from *Plasmodium falciparum* is close to this value (8.1%), suggesting that the positive-inside rule may hold also for this organism.

The fact that the Lys and Arg residues are enriched near the ends of long loops suggests that the positive-inside bias is not confined to short cytoplasmic loops, but rather to regions bordering the transmembrane helices. An outside-biased distribution of (Trp+Tyr) has been noted previously.[38,39] The reason for this bias is unclear, however. Trp and Tyr have a high affinity for the membrane–water interface,[40] although they do not appear to act as topological determinants[41].

In conclusion, our results strongly support the idea that the positive-inside rule is indeed ubiquitous and may be applied to topology predictions on any novel genome. Presumably, the molecular basis for the positive-inside rule is common to all organisms, although it is not yet clear to what extent it depends on properties inherent to the translocon itself[42] or to more general properties of the membrane such as lipid composition[43] and membrane potential.[44,45]

ACKNOWLEDGMENTS

This work was supported by grants from the Swedish Research Council, the Marianne and Marcus Wallenberg Foundation, and the Swedish Cancer Foundation to GvH.

REFERENCES

1. Wallin E, von Heijne G. Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. Protein Sci 1998;7:1029–1038.
2. Jones DT. Do transmembrane protein superfolds exist? FEBS Lett 1998;423:281–285.
3. Sonnhammer EL, von Heijne G, Krogh A. A hidden Markov model for predicting transmembrane helices in protein sequences. In: Glasgow J, editor. Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology. Menlo Park: AAAI Press; 1998. p 175–182.
4. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 2001;305:567–580.
5. Arai M, Ikeda M, Shimizu T. Comprehensive analysis of transmembrane topologies in prokaryotic genomes. Gene 2003;304:77–86.
6. Chen CP, Rost B. State-of-the-art in membrane protein prediction. Appl Bioinformatics 2002;1:21–35.
7. von Heijne G. Control of topology and mode of assembly of a polytopic membrane protein by positively charged residues. Nature 1989;341:456–458.
8. von Heijne G. The distribution of positively charged residues in bacterial inner membrane proteins correlates with the transmembrane topology. EMBO J 1986;5:3021–3027.
9. Sipos L, von Heijne G. Predicting the topology of eukaryotic membrane proteins. Eur J Biochem 1993;213:1333–1340.
10. Gavel Y, Steppuhn J, Herrmann R, von Heijne G. The 'positive-inside rule' applies to thylakoid membrane proteins. FEBS Lett 1991;282:41–46.
11. Gavel Y, von Heijne G. 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 1992;205:1207–1215.
12. von Heijne G. Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. J Mol Biol 1992;225:487–494.
13. Claros MG, von Heijne G. TopPred II: an improved software for membrane protein structure predictions. Comput Appl Biosci 1994;10:685–686.
14. Kessel A, Ben-Tal N. Free energy determinants of peptide association with lipid bilayers. In: Simon S, McIntosh T, editors. Peptide-lipid interactions. San Diego: Academic Press; 2002.
15. Eisenberg D, Schwarz E, Komaromy M, Wall R. Analysis of membrane and surface protein sequences with the hydrophobic moment plot. J Mol Biol 1984;179:125–142.
16. Engelman DM, Steitz TA, Goldman A. Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. Annu Rev Biophys Biophys Chem 1986;15:321–353.
17. Hopp TP, Woods KR. Prediction of protein antigenic determinants from amino acid sequences. Proc Natl Acad Sci USA 1981;78:3824–3828.
18. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol 1982;157:105–132.
19. Jayasinghe S, Hristova K, White SH. Energetics, stability, and prediction of transmembrane helices. J Mol Biol 2001;312:927–934.

20. Argos P, Rao JK. Prediction of protein structure. Methods Enzymol 1986;130:185–207.
21. Zhou H, Zhou Y. Predicting the topology of transmembrane helical proteins using mean burial propensity and a hidden-Markov-model-based method. Protein Sci 2003;12:1547–1555.
22. Möller S, Kriventseva EV, Apweiler R. A collection of well characterised integral membrane proteins. Bioinformatics 2000;16: 1159–1160.
23. Jayasinghe S, Hristova K, White S. A database of membrane protein topology. Protein Sci 2001;10:455–458.
24. Ikeda M, Arai M, Okuno T, Shimizu T. TMPDB: a database of experimentally-characterized transmembrane topologies. Nucleic Acids Res 2003;31:406–409.
25. Westbrook J, Feng Z, Chen L, Yang H, Berman HM. The Protein Data Bank and structural genomics. Nucleic Acids Res 2003;31: 489–491.
26. Carter P, Andersen CA, Rost B. DSSPcont: Continuous secondary structure assignments for proteins. Nucleic Acids Res 2003;31: 3293–3295.
27. Sayle RA, Milner-White EJ. RASMOL: biomolecular graphics for all. Trends Biochem Sci 1995;20:374.
28. Hobohm U, Scharf M, Schneider R, Sander C. Selection of representative protein data sets. Protein Sci 1992;1:409–417.
29. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994;22:4673–4680.
30. Eyrich VA, Marti-Renom MA, Przybylski D, Madhusudhan MS, Fiser A, Pazos F, Valencia A, Sali A, Rost B. EVA: continuous automatic evaluation of protein structure prediction servers. Bioinformatics 2001;17:1242–1243.
31. van den Berg B, Clemons WMJ, Collinson I, Modis Y, Hartmann E, Harrison SC, Rapoport TA. X-ray structure of a protein-conducting channel. Nature 2004;427:36–44.
32. von Heijne G. Recent advances in the understanding of membrane protein assembly and structure. Quart Rev Biophys 2000;32:285–307.
33. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, et al. The COG database: an updated version includes eukaryotes. BMC Bioinformatics 2003;4:41.
34. Hartmann E, Rapoport TA, Lodish HF. Predicting the orientation of eukaryotic membrane-spanning proteins. Proc Natl Acad Sci USA 1989;86:5786–5790.
35. Wallin E, von Heijne G. Properties of N-terminal tails in G-protein coupled receptors: a statistical study. Protein Eng 1995;8:693–698.
36. Juretic D, Zoranic L, Zucic D. Basic charge clusters and predictions of membrane protein topology. J Chem Inf Comput Sci 2002;42:620–632.
37. von Heijne G, Gavel Y. Topogenic signals in integral membrane proteins. Eur J Biochem 1988;174:671–678.
38. Nakashima H, Nishikawa K. The amino acid composition is different between the cytoplasmic and extracellular sides in membrane proteins. FEBS Lett 1992;303:141–146.
39. Schiffer M, Chang CH, Stevens FJ. The functions of tryptophan residues in membrane proteins. Protein Eng 1992;5:213–214.
40. Wimley WC, White SH. Experimentally determined hydrophobicity scale for proteins at membrane interfaces. Nat Struct Biol 1996;3:842–848.
41. Ridder AN, Morein S, Stam JG, Kuhn A, de Kruijff B, Killian JA. Analysis of the role of interfacial tryptophan residues in controlling the topology of membrane proteins. Biochemistry 2000;39: 6521–6528.
42. Goder V, Junne T, Spiess M. Sec61p contributes to signal sequence orientation according to the positive-inside rule. Mol Biol Cell 2004;15:1470–1478.
43. van Klompenburg W, Nilsson IM, von Heijne G, de Kruijff B. Anionic phosholipids are determinants of membrane protein topology. EMBO J 1997;16:4261–4266.
44. Andersson H, von Heijne G. Membrane protein topology: effects of delta mu H+ on the translocation of charged residues explain the 'positive inside' rule. EMBO J 1994;13:2267–2272.
45. van de Vossenberg JL, Albers SV, van der Does C, Driessen AJ, van Klompenburg W. The positive inside rule is not determined by the polarity of the delta psi (transmembrane electrical potential). Mol Microbiol 1998;29:1125–1127.