Protein Science (2000), *9*:505–511. Cambridge University Press. Printed in the USA.
Copyright © 2000 The Protein Society

The effect of nucleotide bias upon the composition and prediction of transmembrane helices

TIMOTHY J. STEVENS AND ISAIAH T. ARKIN

Cambridge Centre for Molecular Recognition, Department of Biochemistry, University of Cambridge, 80 Tennis Court Road, Cambridge CB2 1GA, United Kingdom

(RECEIVED October 13, 1999; FINAL REVISION December 20, 1999; ACCEPTED December 20, 1999)

Abstract

Transmembrane helices are the most readily predictable secondary structure components of proteins. They can be predicted to a high degree of accuracy in a variety of ways. Many of these methods compare new sequence data with the sequence characteristics of known transmembrane domains. However, the known transmembrane sequences are not necessarily representative of a particular organism. We attempt to demonstrate that parameters optimized for the known transmembrane domains are far from optimal when predicting transmembrane regions in a given genome. In particular, we have tested the effect of nucleotide bias upon the composition and hence the prediction characteristics of transmembrane helices. Our analysis shows that nucleotide bias of a genome has a strong and predictable influence upon the occurrences of several of the most important hydrophobic amino acids found within transmembrane helices. Thus, we show that nucleotide bias should be taken into account when determining putative transmembrane domains from sequence data.

Keywords: genome; hydropathy; membrane protein; proteome

Membrane proteins are of extreme importance in biomedical research. Most pharmaceuticals in use today interact with transmembrane proteins. Membrane proteins are also extremely abundant, and previous analyses (Arkin et al., 1997) suggest that about one-quarter of all proteins are anchored in a lipid bilayer by virtue of a transmembrane α-helix. Given this abundance, the number of potentially therapeutic targets is huge. It is only now, with the advent of mass genome sequencing projects, that we can begin to realize the whole scope of potential transmembrane targets, with the first step along this route being the prediction of transmembrane spans from sequence data.

Membrane α-helices are generally characterized by a hydrophobic stretch of between 12 and 25 amino acids (von Heijne, 1995; Whitley et al., 1996). The amino acid composition of these transmembrane regions has given rise to the many hydrophobicity (Kyte & Doolittle, 1982; Engelman et al., 1986; Rees et al., 1989; Degli Esposito et al., 1990; White & Wimley, 1999) and membrane preference scales (von Heijne, 1992; Li & Deber, 1994a). Searching new sequences using these membrane preference parameters allows quick and accurate domain prediction. Indeed, this is the most accurate protein domain classification known, which is based upon sequence data alone. More recently, even better methods have appeared, most notably those that employ neural network algorithms (Rost et al., 1995, 1996). These methods learn the characteristics of the known transmembrane regions during a training period, after which they can predict the presence or absence of transmembrane regions in new sequences. Both the membrane preference scales and the neural network methods use the known transmembrane regions as their foundation. As the known transmembrane domains are severely biased toward model organisms and higher eukaryotes, we have decided to investigate how different hydropathy analysis parameters predict transmembrane domains in different genomes. We wish to test whether a single set of parameters, optimized on the known transmembrane helices can be applied, with equal results, to a diverse range of different organisms. If the amino acid composition of transmembrane domains varies significantly according to organism, then a single set of search parameters, derived from the known helices, will not take this into account. Hence, we aim to determine if a readily obtainable organism specific parameter, genomic nucleotide bias, affects the amino acid composition of transmembrane domains.

There are now a number of completed genomes, with representatives from all of the three domains of life. Our analysis uses proteomic databases constructed from the predicted open reading frames of these genomes. In each proteome, we have determined the compositions of the putative transmembrane helices and the

Reprint requests to: Isaiah T. Arkin, Cambridge Centre for Molecular Recognition, Department of Biochemistry, University of Cambridge, 80 Tennis Court Road, Cambridge CB2 1GA, United Kingdom; e-mail: sa232@cam.ac.uk.

Abbreviations: AT, deoxyadenylic acid-thymidylic acid; GC, deoxyguanylic acid-deoxycylidylic acid; ORF, open reading frame; TM, transmembrane.

Table 1. *A table to show the nucleotide bias in the genomes of the organisms studied*

Organism	%GC/%AT
B. burgdorferi	0.400
R. prowazekii	0.408
M. jannaschii	0.458
M. genitalium	0.464
H. influenzae	0.617
H. pylori	0.636
S. cerevisiae	0.656
M. pneumoniae	0.664
C. pneumoniae	0.683
C. trachomatis	0.704
P. horikoshii	0.721
C. elegans	0.742
A. aeolicus	0.769
B. subtilis	0.770
S. PCC6803	0.913
A. fulgidus	0.945
M. thermoautotrophicum	0.982
E. coli	1.032
T. pallidum	1.118
M. tuberculosis	1.908

number of putative transmembrane proteins predicted by different hydropathy search parameters. Our analyses reveal that the number of putative transmembrane domains, predicted across a wide range of search parameters, varies according to the organism, and that the composition of transmembrane domains is different in different organisms. We illustrate that the differing composition of the transmembrane domains is strongly correlated with, and may be due, at least in part, to the nucleotide bias of the organism's genome.

Results

Nucleotide bias

The nucleotide biases in each of the organisms are illustrated in Table 1. This illustrates that there is significant variation in the nucleotides bias within the genomes studied. The codon use of the abundant (Deber et al., 1986) transmembrane amino acids is illustrated in Table 2. Where an amino acid is only coded by AT or GC rich codons, a significant AT/GC bias within a genome is expected to constrain the abundance of the relevant amino acids, provided alternative residues can be used.

Table 2. *A table to show the codons for the abundant transmembrane amino acids and the minimum number of AT and GC bases required to code for the amino acid*

Amino acid	Codons	Min. A + T	Min. G + C
Ala	GCX	0	2
Gly	GGX	0	2
Val	GTX	1	1
Leu	CTX TTA/G	1	1
Ile	ATA/C/T	2	0
Phe	TTC/T	2	0

Amino acid composition

The plot presented in Figure 1 depicts the composition of the TM regions, from each of the proteomes in our database, for which we can be most certain of a transmembrane identity. The data illustrated (Fig. 1) represents 409,890 residues in 26,436 predicted TM helices from 8,269 genes. Of the abundant, hydrophobic amino acids leucine has a high abundance in all organisms. Isoleucine and phenylalanine, however, have a high abundance in some organisms and a low abundance in others. From this chart, it is apparent that when isoleucine and phenylalanine are of low abundance, valine and alanine abundance is higher and vice versa. Although valine does not necessarily have an AT or GC rich codon, it would seem that its abundance is elevated when the genome is biased against AT bases. This suggests that the abundance of valine is compensating for the lack of the AT constrained amino acids, particularly isoleucine, with which it shares similar physical properties. Also of note is that isoleucine complement is not particularly substituted by leucine, which is the most abundant transmembrane residue. The standard deviation in the abundance of each amino acid across the different organisms, as presented in Table 3, confirms that the amino acids Ile, Phe, Val, and Ala have the greatest variation in proteomic abundance.

The graph illustrated in Figure 2 shows further conformation of the correlated abundances of the amino acids Ile, Phe, Val, and Ala. The Phe + Ile composition (FI) has a linear relationship with the Val + Ala composition (VA). A lower proportional composition of Phe and Ile in a proteome, and within its TM domains, is compensated for by a higher proportion of Val and Ala and vice versa. Leucine did not show any significant correlation in its abundance with any other amino acid (data not shown). The graph presented

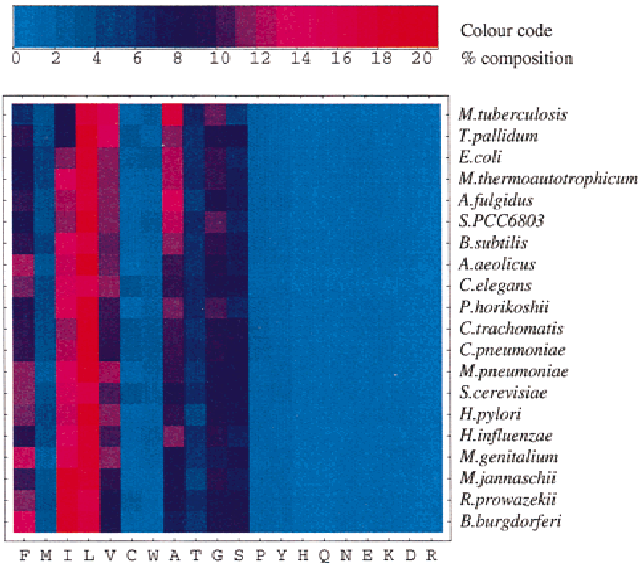


Fig. 1. A plot to show the amino acid composition of the TM domains in each of the proteomes under investigation. Percentage of TM composition exhibited by a residue in each organism increases from blue to red.

Table 3. *A table to show the mean, μ, and standard deviation, σ, for the % abundances of each amino acid across the organisms studied both within TM domains (μ_{TM}, σ_{TM}) and for the complete protein sequence (μ_C, σ_C)*

Amino acid	σ_{TM}	μ_{TM}	σ_C	μ_C
F	1.86	10.67	0.79	4.69
M	0.66	3.83	0.40	2.23
I	3.16	14.48	1.76	7.37
L	1.26	19.53	0.74	10.25
V	1.70	11.64	1.00	6.80
C	0.96	1.57	0.40	1.15
W	0.48	1.81	0.27	1.03
A	2.44	11.50	1.98	7.31
T	0.54	5.89	0.65	5.09
G	1.13	9.68	1.23	6.50
S	1.05	4.63	1.16	6.50
P	0.21	1.93	0.78	4.03
Y	0.26	1.55	0.54	3.39
H	0.06	0.23	0.40	1.96
Q	0.09	0.34	1.26	3.48
N	0.15	0.53	1.53	4.63
E	0.04	0.09	1.28	6.84
K	0.05	0.06	2.29	6.91
D	0.02	0.05	0.47	5.01
R	0.01	0.00	1.30	4.82

in Figure 3 confirms that alternate hydrophobic amino acid use within TM domains (FI vs. VA) is strongly correlated with the nucleotide bias of the genome (GC vs. AT). Indeed, it would appear that, over the range studied, FI/VA is proportional to GC/AT, with a gradient of 0.827 (correlation coefficient = 0.987). From this it is clear that nucleotide bias imposes constraints upon codon use, and hence, amino acid composition. As the hydrophobicity of isoleucine and phenylalanine is greater than the compositionally compensating TM domain amino acids valine and alanine, one may also expect that when GC bias favors valine–alanine use, the average hydrophobicity of the TM α-helix residues decreases.

Hydropathy searches

The plots presented in Figure 4 illustrates how the predicted % of membrane proteins in each of the proteomes varies with the simple hydropathy search parameters of window size and hydrophobicity threshold. By ordering these plots according to decreasing AT:GC bias (%GC/%AT), it can clearly be seen that the pattern of prediction varies. The regions in Figure 4 that correspond to permissive hydropathy searches (blue areas) do not vary markedly across organisms. However, this is expected as a limitation on hydrophobicity, imparted by AT:GC bias, will only effect hydropathy searches that use restrictive search parameters (red region). The transmembrane predictions that arise from permissive searches that are dom-

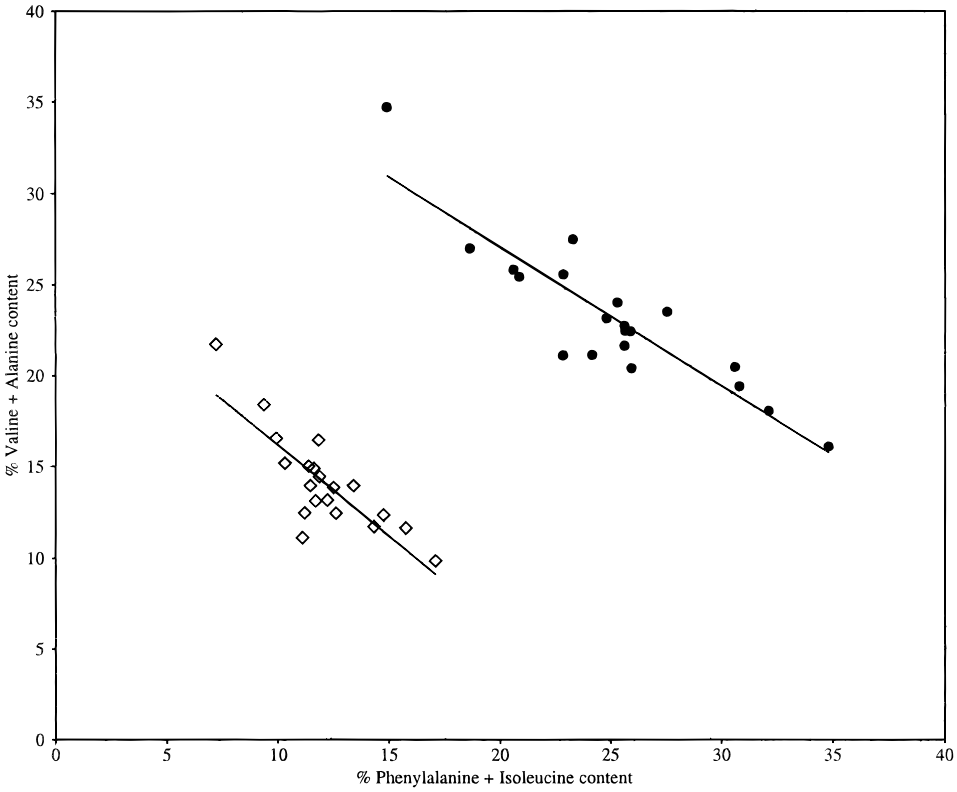


Fig. 2. A graph to show the correlation between the percentage abundance of valine plus alanine residues (VA) and the percentage abundance of phenylalanine plus isoleucine residues (FI). Data are shown for each organism, for both the amino acids within the predicted TM domains (●) and the whole proteome (◇).

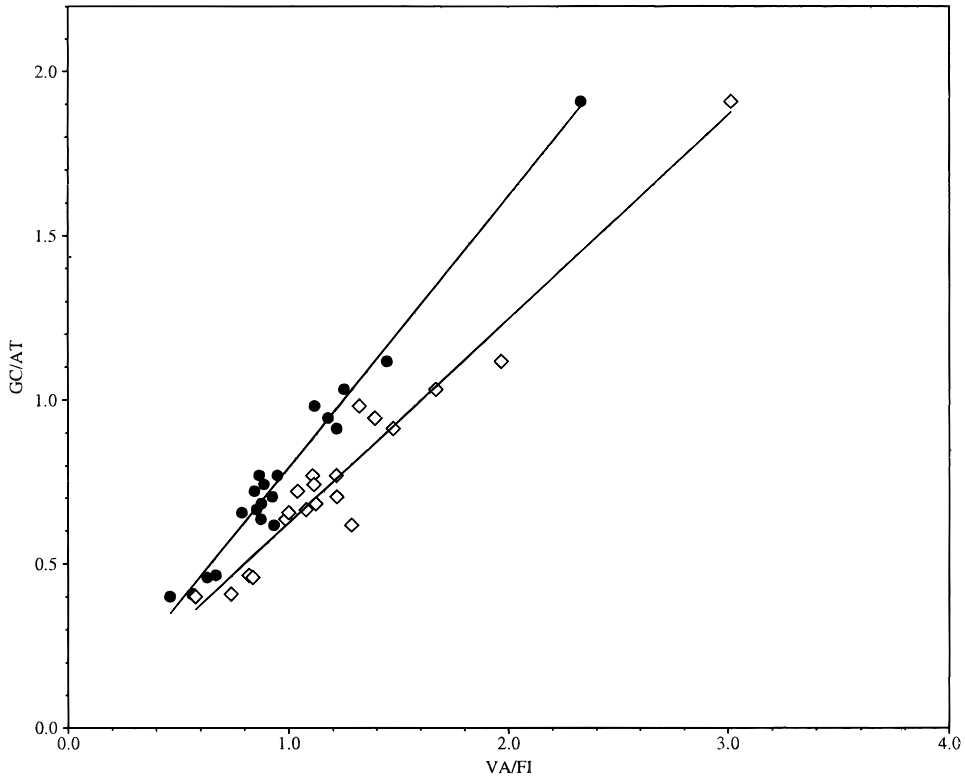


Fig. 3. A graph to show the correlation between the nucleotide bias of a genome (GC/AT) and alternate hydrophobic amino acid use (VA/FI). Data are shown for both the amino acids within the predicted TM domains (●) and for the whole proteome (◇).

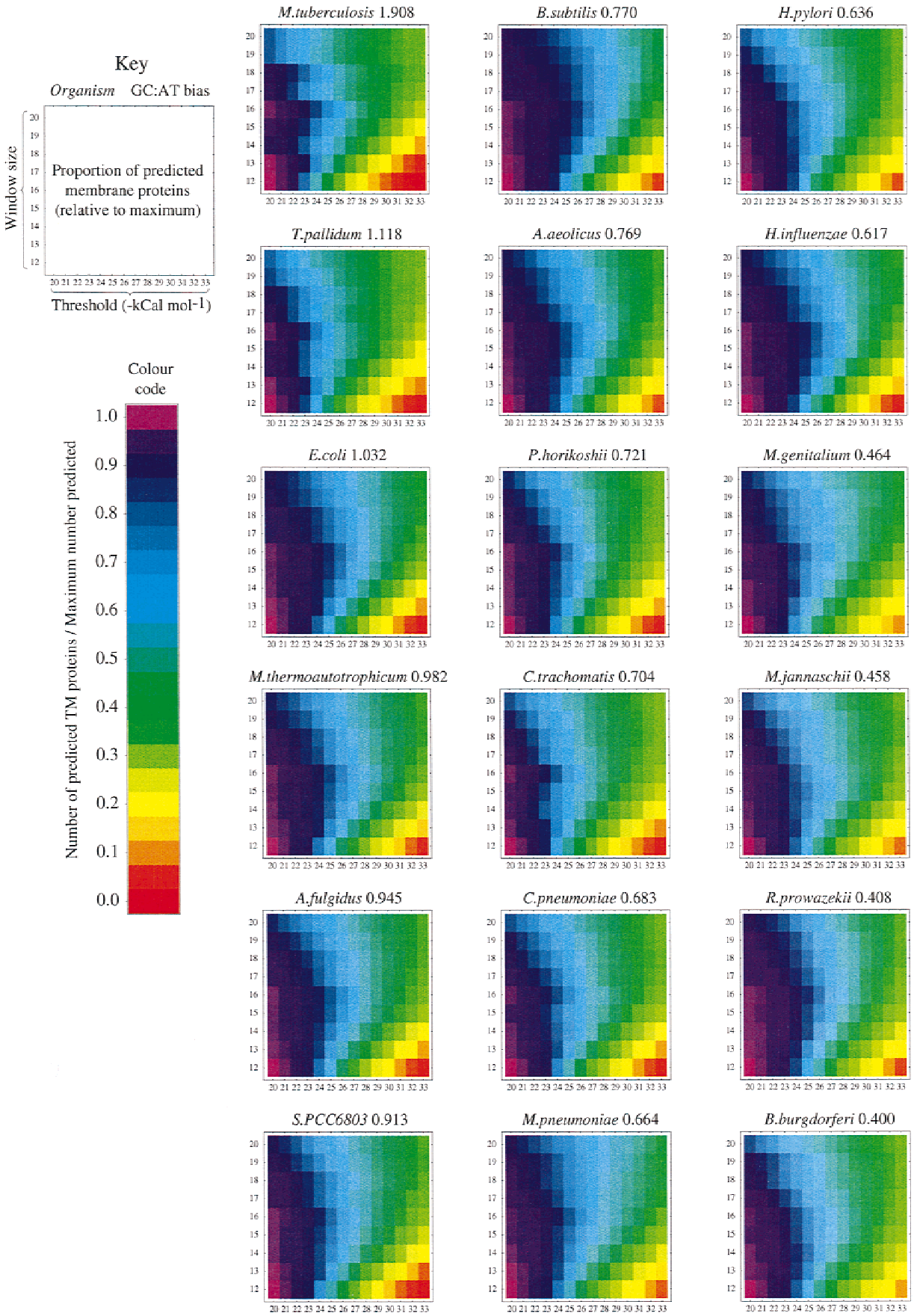
inated by false positive results and, hence, domains of reduced hydrophobicity are not resolvable.

We cannot directly measure the length and, hence, the average residue hydrophobicity of the transmembrane regions identified. This is because of difficulties encountered in defining the ends of the transmembrane helices. The analysis we present here uses a consistent but arbitrary means of determining the ends of the transmembrane domains, based upon a hydrophobicity. Although we are satisfied that this is sufficient to determine the bulk composition of TM domains across a whole proteome, the ends are not sufficiently well defined to make confident length and mean hydrophobicity measurements. However, although we cannot accurately determine average hydrophobicity, it is clear from the multiparameter hydropathy search plots (Fig. 4) that stringent hydropathy searches (small window, large threshold) predict fewer membrane proteins in genomes with restricted AT content. Thus, assuming that the proportion of membrane proteins in a proteome does not correlate with nucleotide bias, restricted AT content gives rise to transmembrane helices with less hydrophobic residues, consistent with the alternate residue content described above. The use of less hydrophobic amino acids within transmembrane domains may give rise to longer transmembrane helices; however, this is not resolvable in this analysis.

Discussion

As the amino acids phenylalanine and isoleucine seem to be so readily substitutable for valine and alanine (Fig. 2) within transmembrane helices, according to the base pair preferences of the organism (Fig. 3), the general, bulk composition of transmembrane α-helices would seem to be able to vary quite dramatically while the function of proteins is preserved (Zhou et al., 1997). Moreover, the correlation between nucleotide bias and alternate amino acid use is so good that we suggest that GC/AT can be used to accurately predict VA/FI and vice versa. For the known membrane spanning sections, as described in the SWISS-PROT database, VA/FI = 1.05. According to the relationship described above, this would correspond to a GC/AT of about 0.87, an intermediate nucleotide bias similar to the eukaryotic genomes studied. Further, as each of the hydropathy "landscapes" (illustrated in Fig. 4) is particular to the organism, the hydropathy search characteristic of the known transmembrane domains cannot be representative of a

Fig. 4 (*facing page*)**.** The amount of predicted membrane proteins using hydropathy thresholds from -20 to -33 kcal mol^{-1} and window sizes from 12 to 20 residues, for each of the organisms studied. The plots are displayed in order of decreasing genome GC/AT bias (given after the organism name). In each plot the color scale indicates the proportion of predicted membrane proteins, relative to the maximum number predicted, at the most permissive parameters for the plot (-20 kcal mol^{-1} threshold and a window of 12).


Key
Organism GC:AT bias
Window size
Proportion of predicted membrane proteins (relative to maximum)
Threshold (-kCal mol-1)
Colour code
Number of predicted TM proteins / Maximum number predicted
M.tuberculosis 1.908
B.subtilis 0.770
H.pylori 0.636
T.pallidum 1.118
A.aeolicus 0.769
H.influenzae 0.617
E.coli 1.032
P.horikoshii 0.721
M.genitalium 0.464
M.thermoautotrophicum 0.982
C.trachomatis 0.704
M.jannaschii 0.458
A.fulgidus 0.945
C.pneumoniae 0.683
R.prowazekii 0.408
S.PCC6803 0.913
M.pneumoniae 0.664
B.burgdorferi 0.400

particular organism and will show marked differences to proteomes that are under the influence of a significant nucleotide bias.

Throughout this analysis, we have use the GES hydrophobicity scale (Engelman et al., 1986). Although there are many alternative hydrophobicity and membrane preference scales used in the prediction of transmembrane domains (Kyte & Doolittle, 1982; Li & Deber, 1994b; White & Wimley, 1999), we do not believe that the use of a particular scale will have a substantial influence upon the results of this analysis. The allocation of transmembrane domains within the proteins is made primarily by homology to known membrane proteins with only the use of an extremely permissive hydropathy search to ensure that the sequences compared in the alignment have some degree of membrane character. When searching proteomes with various search parameters, different scales will have different prediction results. However, we believe that the basic result of this analysis remains, in that the composition and, hence, predictability of transmembrane domains seems to vary according to nucleotide bias. Thus, for any fixed membrane-preference scale, the best search parameters will vary according to organism. It is noted that in some hydrophobicity scales (White & Wimley, 1999), the difference in the hydrophobicity between Val + Ala and Phe + Ile is greater than in the GES scale used here; hence, with these scales the average hydrophobicity of a membrane residue may vary even more markedly according to the organism.

Although the analysis we present is mostly based upon data derived from prokaryotic sequences, the results may be of significance for the analysis of eukaryotic genomes. Within eukaryotic genomes there are different regions of DNA which have distinct, but consistent nucleotide biases (Sabeur et al., 1993; Bradnam et al., 1999). The particular GC:AT bias of these isochore regions may have a significant and predictable effect upon the composition of transmembrane domains, analogous to the results presented above. Thus, it may be possible to improve the prediction of the transmembrane domains for eukaryotic sequences if the GC:AT bias of the isochore, in which the gene lies, is known. However, further analyses will be required to confirm an icochore specific influence on eukaryotic transmembrane composition. Indeed, it may be of value to also consider the local GC:AT context within prokaryotic genomes, as there are some regions of distinct nucleotide bias, often attributed to insertion events (Garcia-Vallve et al., 1999).

Conclusion

We have shown that in organisms where the nucleotide content of the genome is significantly biased, the composition of its transmembrane helices is altered. A genomic bias against AT or GC bases constrains codon use and, hence, may constrain use of certain hydrophobic amino acids. A relatively low abundance of these residues leads to an increase in the use of alternative amino acids, with a more favorable codon. However, as these substitute residues do not have the same hydrophobicity, the overall hydrophobicity of TM domains is altered. Specifically, a bias against AT bases leads to less hydrophobic transmembrane residues. As the known membrane domains present an average picture, TM prediction analyses, based upon this or any other nonorganism specific data set alone, will be ignorant of the significant compositional differences that exist in different proteomes. However, as we have shown, VA/FI is readily predictable; GC/AT $\approx$ 0.83(VA/FI). Thus, AT bias can and should be taken into account when undertaking transmembrane prediction on a sequence from a known organism.

Materials and methods

Database selection

A proteome database was constructed, which consists of the protein sequences of predicted ORFs from the organisms that, to this date, have had their genomes completely sequenced. These organisms include members of archea: *Archaeoglobus fulgidus* (Klenk et al., 1997), *Methanobacterium thermoautotrophicum* (Smith et al., 1997), *Methanococcus jannaschii* (Bult et al., 1996), and *Pyrococcus horikoshii* (Kawarabayasi et al., 1998); eubacteria: *Aquifex aeolicus* (Deckert et al., 1998), *Bacillus subtilis* (Kunst et al., 1997), *Borrelia burgdorferi* (Fraser et al., 1997), *Chlamydia pneumoniae* (Kalman et al., 1999), *Chlamydia trachomatis* (Stephens et al., 1998), *Escherichia coli* (Blattner et al., 1997), *Haemophilus influenzae* (Fleischmann et al., 1995), *Helicobacter pylori* (White et al., 1997), *Mycobacterium tuberculosis* (Cole et al., 1998), *Mycoplasma genitalium* (Fraser et al., 1995), *Mycoplasma pneumoniae* (Himmelreich et al., 1996), *Rickettsia prowazekii* (Andersson et al., 1998), *Synechocystis* PCC6803 (Nakamura et al., 1998), and *Treponema pallidum* (Fraser et al., 1998); and eukaryota: *Caenorhabditis elegans* (CESC, 1998), and *Saccharomyces cerevisiae* (Chervitz et al., 1999). Initially, for each genome the nucleotide bias was calculated; the ratio between the combined abundance of C and G bases and the abundance of the A and T bases.

Transmembrane composition

We have predicted the occurrence of transmembrane helices within our proteome database on the basis of both homology to known transmembrane proteins and hydropathy analysis. The set of known transmembrane proteins used in this analysis are those proteins within SWISS-PROT (Bairoch & Boeckmann, 1991), which have a predicted hydrophobic membrane spanning region ("TRANSMEM" within the FT field). Each of these (11,110) protein sequences was compared to all of the ORFs within for each of our organisms using the FASTA (Pearson & Lipman, 1988) algorithm. In each instance, the best matching protein within a given proteome was selected for further analysis if the random probability of a homology match was $<10^{-5}$. A FASTA alignment of each SWISS-PROT membrane protein and corresponding proteome homologue was obtained. Simple hydropathy searches of the aligned sequences were performed, which considered ver permissive parameters; a window of 12 residues and a hydrophobicity threshold of -22 kcal mol^{-1}. Where the hydropathy searches identified hydrophobic sections in both sequences that overlap with each other and the SWISS-PROT assignment of the putative transmembrane region (according to the FASTA alignment), the hydrophobic region of the proteomic sequence was assigned as transmembrane. The amino acid composition, length, and average residue hydrophobicity of the predicted transmembrane helix was then calculated and tabulated. Although this method will not identify all of the transmembrane spans within the proteomes, the regions have not been identified solely on the basis of an amino acid preference scale. Also, by using a correlation between hydrophobic regions in a homologue and a SWISS-PROT transmembrane assignment, we attempt to minimize false-positive classification.

Hydropathy searches

To show how the transmembrane α-helix prediction results vary between organisms, hydropathy searches were performed upon

each proteome using a wide range of search parameters. The hydrophobicity searches consider a given window of residues within an amino acid sequence and classify a region as transmembrane if the total GES scale (Engelman et al., 1986) hydrophobicity exceeds a particular threshold value. We searched each of the proteomes using window sizes ranging from 12 to 26 residues and all the integer threshold values, ranging from -15 to -40 kcal mol^{-1}. The resulting number of putative membrane proteins predicted for each combination of parameters was recorded for each organism.

Acknowledgments

This work was supported by grants from the Wellcome trust and the Biotechnology and Biological Sciences Research Council to I.T.A.

References

Andersson SGE, Zomorodipour A, Andersson JO, SicheritzPonten T, Alsmark UCM, Podowski RM, Naslund AK, Eriksson AS, Winkler HH, Kurland CG. 1998. The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature 396*:133–140.

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

Bairoch A, Boeckmann B. 1991. The SWISS-PROT protein-sequence databank. *Nucleic Acids Res 19*:2247–2248.

Blattner FR, Plunkett G III, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, et al. 1997. The complete genome sequence of *Escherichia coli* K-12. *Science 277*:1453–1462.

Bradnam KR, Seoighe C, Sharp PM, Wolfe KH. 1999. G+C content variation along and among *Saccharomyces cerevisiae* chromosomes. *Mol Biol Evol 16*:666–675.

Bult CJ, White O, Olsen GJ, Zhou LX, Fleischmann RD, Sutton GG, Blake JA, FitzGerald LM, Clayton RA, Gocayne JD, et al. 1996. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science 282*:1058–1073.

CESC (*C. elegans* Sequencing Consortium). 1998. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science 229*: 2012–2018.

Chervitz SA, Hester ET, Ball CA, Dolinski K, Dwight SS, Harris MA, Juvik G, Malekian A, Roberts S, Roe T, et al. 1999. Using the *Saccharomyces* Genome Database (SGD) for analysis of protein similarities and structure. *Nucleic Acid Res 27*:74–78.

Cole ST, Brosch R, Parkhill J, Gamier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE, et al. 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature 393*:537–544.

Deber CM, Brandl CJ, Deber RB, Hsu LC, Young XK. 1986. Amino acid composition of the membrane and aqueous domains of integral membrane proteins. *Arch Biochem Biophys 251*:68–76.

Deckert G, Warren PV, Gaasterland T, Young WG, Lenox AL, Grahams DE, Overbeek R, Snead MA, Keller M, Aujay M, et al. 1998. The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. *Nature 392*:353–358.

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

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, et al. 1995. Whose-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science 269*:496–521.

Fraser CM, Casjens S, Huang WM, Sutton GG, Clayton R, Lathigra R, White O, Ketchum KA, Dodson R, Hickey EK, et al. 1997. Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. *Nature 390*:580–586.

Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ, Kerlavage AR, Sutton G, Kelley JM, et al. 1995. The minimal gene complement of *Mycoplasma genitalium*. *Science 270*:397–403.

Fraser CM, Norris SJ, Weinstock GM, White O, Sutton GG, Dodson R, Gwinn M, Hickey EK, Clayton R, Ketchum KA, et al. 1998. Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. *Science 281*:375–388.

Garcia-Vallve S, Palau J, Romeu A. 1999. Horizontal gene transfer in glycosyl hydrolases inferred from codon usage in *Escherichia coli* and *Bacillus subtilis*. *Mol Biol Evol 16*:1125–1134.

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 Acid Res 24*:4420–4449.

Kalman S, Mitchell W, Marathe R, Lammel C, Fan L, Hyman RW, Olinger L, Grimwood L, Davis RW, Stephens RS. 1999. Comparative genomes of *Chlamydia pneumoniae* and *C-trachomatis*. *Nat Genet 21*:385–389.

Kawarabayasi Y, Sawada M, Horikawa H, Haikawa Y, Hino Y, Yamamoto S, Sekine M, Baba S, Kosugi H, Hosoyama A, et al. 1998. Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium, *Pyrococcus horikoshii* 0T3. *DNA Res 5*:55–76.

Klenk HP, Clayton RA, Tomb JF, White O, Nelson KE, Ketchum KA, Dodson RJ, Gwinn M, Hickey EK, Peterson JD, et al. 1997. The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. *Nature 390*:364–370.

Kunst F, Ogasawara N, Moszer I, Albertini AM, Alloni G, Azevedo V, Bertero MG, Bessieres P, Bolotin A, Borchert S, et al. 1997. The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature 390*: 249–256.

Kyte J, Doolittle RF. 1982. A simple method for displaying the hydropathic character of a protein. *J Mol Biol 157*:105–132.

Li SC, Deber CM. 1994a. A measure of helical propensity for amino acids in membrane environments. *Nat Struct Biol 1*:558.

Li SC, Deber CM. 1994b. A measure of helical propensity for amino acids in membrane environments. *Nat Struct Biol 1*:368–373.

Nakamura Y, Kaneko T, Hirosawa M, Miyajima N, Tabata S. 1998. CyanoBase, a www database containing the complete nucleotide sequence of the genome of *Synechocystis* sp. strain PCC6803. *Nucleic Acid Res 26*:63–67.

Pearson WR, Lipman DJ. 1988. Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA 85*:2444–2448.

Rees DC, DeAntonio L, Eisenberg D. 1989. Hydrophobic organization of membrane proteins. *Science 245*:510–513.

Rost B, Casadio R, Fariselli P, Sander C. 1995. Transmembrane helices predicted at 95-percent accuracy. *Protein Sci 4*:521–533.

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

Sabeur G, Macaya G, Kadi F, Bernardi G. 1993. The isochore patterns of mammalian genomes and their phylogenetic implications. *J Mol Evol 37*:93–108.

Smith DR, Doucette-Stamm LA, Deloughery C, Lee HM, Dubois J, Aldredge T, Bashirzadeh R, Blakely D, Cook R, Gilbert K, et al. 1997. Complete genome sequence of *Methanobacterium thermoautotrophicum* DeltaH: Functional analysis and comparative genomics. *J Bacteriol 179*:7135–7155.

Stephens RS, Kalman S, Lammel C, Fan J, Marathe R, Aravind L, Mitchel W, Olinger L, Tatusov RL, Zhao Q, et al. 1998. Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. *Science 282*: 754–759.

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

von Heijne G. 1995. Membrane-protein assembly—Rules of the game. *Bioessays 17*:25–30.

White O, Kerlavage AR, Clayton RA, Sutton GG, Fleischmann RD, Ketchum KA, Klenk HP, Gill S, Dougherty BA, Nelson K, et al. 1997. The complete genome sequence of the gastric pathogen *Helicobacter pylon*. *Nature 388*:539–547.

White SH, Wimley WC. 1999. Membrane protein folding and stability: Physical principles. *Annu Rev Biophys Biomol Struct 28*:319–365.

Whitley P, Grahn E, Kutay IT, Rapoport TA, von Heijne G. 1996. A 12-residue-long polyleucine tail is sufficient to anchor synaptobrevin to the endoplasmic reticulum membrane. *J Biol Chem 271*:7583–7586.

Zhou YF, Wen J, Bowie JU. 1997. A passive transmembrane helix. *Nat Struct Biol 4*:986–990.