PROTEINS: Structure, Function, and Genetics 39:417–420 (2000)

Do More Complex Organisms Have a Greater Proportion of Membrane Proteins in Their Genomes?

Timothy J. Stevens and Isaiah T. Arkin*
Cambridge Centre for Molecular Recognition, Department of Biochemistry, University of Cambridge, Cambridge, United Kingdom

ABSTRACT **One may speculate that higher organisms require a proportionately greater abundance of membrane proteins within their genomes in order to furnish the requirements of differentiated cell types, compartmentalization, and intercellular signalling. With the recent availability of several complete prokaryotic genome sequences and sufficient progress in many eukaryotic genome sequencing projects, we seek to test this hypothesis. Using optimized hydropathy analysis of proteins in several, diverse proteomes, we show that organisms of the three domains of life—Eukarya, Eubacteria, and Archaea—have similar proportions of α-helical membrane proteins within their genomes and that these are matched by the complexity of the aqueous components. Proteins 2000;39:417–420.**
© 2000 Wiley-Liss, Inc.

Key words: proteome; protein sequence; α-helix; hydrophobicity; hydropathy

INTRODUCTION

There is no doubt that membrane proteins are extremely important in biomedicine; indeed they are the targets for most of the pharmaceuticals in use today. So that we can realize the scope of transmembrane proteins, a pertinent question would be to ask how many membrane proteins there are. Also, with knowledge of the occurrences of membrane proteins across a wide range of organisms, one can begin to see if there is any correlation between the abundance of membrane proteins and the characteristics of a given class of organisms. Previous studies have suggested that there is, and that the proportion of membrane proteins increases as a function of genome size.[5] However, we believe that is not so and that the data must be re-examined.

Eukaryotic organisms, by definition, are more complex than the organisms in the kingdoms of Eubacteria and Archaea. Eukaryotes, unlike the majority of prokaryotes, are required to regulate the passage of substances through the internal membranes that define their intracellular compartments. Also, metazoan eukaryotes may have a different pattern of expression for each cell type and so have developmentally specific membrane requirements and a greater need for intercellular communication. One might expect eukaryotes to devote a greater proportion of their genomes to membrane proteins in order to support these membrane processes. Alternatively, more complex organisms, with more proteins, may not have a larger complement of membrane proteins, if the increase in functionality at membranes is matched by an increase in the complement of aqueous components. To test this, optimized hydropathy searches were performed upon proteome sequences from several prokaryotic (all complete) and several eukaryotic (mostly incomplete) databases. Where genome data was incomplete, the non-homologous sequences studied were of both expressed sequence tag (EST) and genomic origin. Here it was necessary to estimate the proteome sizes, mostly based upon the number of proteins identified in these incomplete genome surveys. For each proteome, the proportion of proteins that contain a putative hydrophobic transmembrane α-helix was calculated.

METHODS

We have used different sets of hydropathy search parameters to give some indication of the confidence bounds associated with the estimation of membrane protein complement from sequence data. The searches of the proteomes studied (Table I) were performed with three sets of parameters: restrictive, optimized, and permissive. These search parameters are derived by performing hydropathy analysis on the sequences of the known membrane proteins in the protein data bank (PDB) database. The restrictive hydropathy searches considered a window of 12 residues with a total GES scale[1] hydropathy of less than -33 kCal mole^{-1}. This represents parameters chosen to give zero false-positive classifications within the Brookhaven PDB. The permissive parameters corresponded to a window of 15 residues and a hydropathy threshold of -22 kCal mole^{-1}, which was chosen to give zero false-negatives within the membrane proteins of the PDB. Last, the optimized search used a window of 15 residues and a threshold of -27 kCal mole^{-1}, so as to give the minimum overall

Grant sponsor: Wellcome Trust; Grant sponsor: Biotechnology and Biological Sciences Research Council.

*Correspondence to: Isaiah T. Arkin, Cambridge Centre for Molecular Recognition, Department of Biology, University of Cambridge, 80 Tennis Court Road, Cambridge, CB2 1GA, United Kingdom. E-mail: sa232@cam.ac.uk

Received 5 October 1999; Accepted 4 February 2000

© 2000 WILEY-LISS, INC.

TABLE I. The Estimated Percent of Membrane Proteins and Proteome Sizes for the Organisms Studied[†]

Organism	Percent membrane proteins	Proteins studied	Size of proteome	Percent genome sequenced
Archea				
Archaeoglobus fulgidus	24.2	2,409	2,409	100
Methanococcus jannaschii	20.4	1,715	1,715	100
Methanobacterium thermoauto.	24.9	1,871	1,871	100
Pyrococcus horikoshii	29.9	2,061	2,061	100
Eubacteria				
Aquifex aeolicus	24.3	1,522	1,522	100
Borellia burgdorferi	29.3	847	847	100
Bacillus subtilis	29.2	4,100	4,100	100
Chlamydia pneumoniae	33.3	1,052	1,052	100
Chlamydia trachomatis	30.0	894	894	100
Escherichia coli	29.9	4,290	4,290	100
Haemophilus influenzae	25.3	1,707	1,707	100
Helicobacter pylori	25.1	1,577	1,577	100
Mycoplasma genitalium	27.2	467	467	100
Mycoplasma pneumoniae	26.4	677	677	100
Mycobacterium tuberculosis	29.9	3,918	3,918	100
Rickettsia prowazekii	31.2	834	834	100
Synechocystis PCC6803	30.6	3,169	3,169	100
Treponema pallidum	29.2	1,030	1,030	100
Eukarya				
Arabidopsis thaliana	30.5	8,919	20,000	51
Caenorhabditis elegans	40.9	14,703	19,000	84
Drosophila melanogaster	24.9	2,697	16,000	25
Homo sapiens	29.7	15,144	70,000	11
Saccharomyces cerevisiae	28.2	6,243	6,243	100
Schizosaccharomyces pombe	23.1	3,927	4,100	85

[†]Note that where genome sequence is incomplete the proteins studied are derived from EST and genomic sources.

error (positive plus negative) when searching the PDB. All searches employed a simple algorithm to avoid the misclassification of N-terminal signal peptide regions.[2,3] Hydrophobic regions were ignored if they occurred within the first 30 amino-terminal amino acids and carried a net positive charge.

The use of EST data in this analysis of most of the Eukaryotic organisms was thought to be useful, given the state of progress in the genome sequence surveys. For example, at this time in the Human Genome Project only about 11% of the total sequence has been completed and made available, but more than half of the estimated total protein complement is known, primarily from EST data. We believe the large number of ESTs and the sensitivity limit of EST generation is now sufficient to give a representative picture of the abundance of membrane proteins, even though, in general, membrane proteins are transcribed at lower levels than aqueous proteins.

RESULTS AND DISCUSSION

Figure 1 illustrates the results of the analyses used to determine the hydropathy search parameters. These plots illustrate that there are a range of values for which the false-positive and false-negative search error is minimized. The particular minima chosen for the permissive and restrictive hydropathy searches are those which are closest to the parameters for the minimum total error. As is shown in Figure 2, it is clear that there is no resolvable correlation between proteome size and the proportion of membrane proteins contained therein. For all three sets of search parameters, there is no significant difference between the different domains of life, and there is no pattern within any domain. The greatest proportion of putative membrane proteins (41%) was found in *C. elegans,* whereas for all other organisms the estimated proportion of membrane proteins is between 20% and 35% (see Table I). Part of the variation in the prediction of helices may be indicative of the search method, rather than variations in membrane protein complement. Any organism-specific differences in the bulk composition of transmembrane domains will mean that a given set of search parameters will show some genome-specific variation in its prediction, even if the search is optimized using the known membrane proteins. However, as there is no correlation between

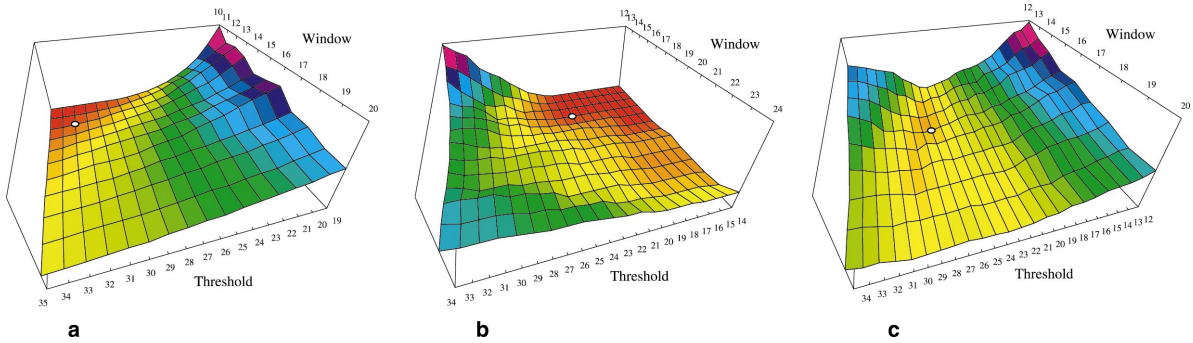

Fig. 1. A figure to illustrate the false-positive (**a**), false-negative (**b**), and total (**c**) assignment errors when performing hydropathy searches of the PDB database for various combinations of window size and hydropathy threshold (units of -kCal mole^{-1}). The white circles indicate the points of minimum error used in subsequent hydropathy analyses.

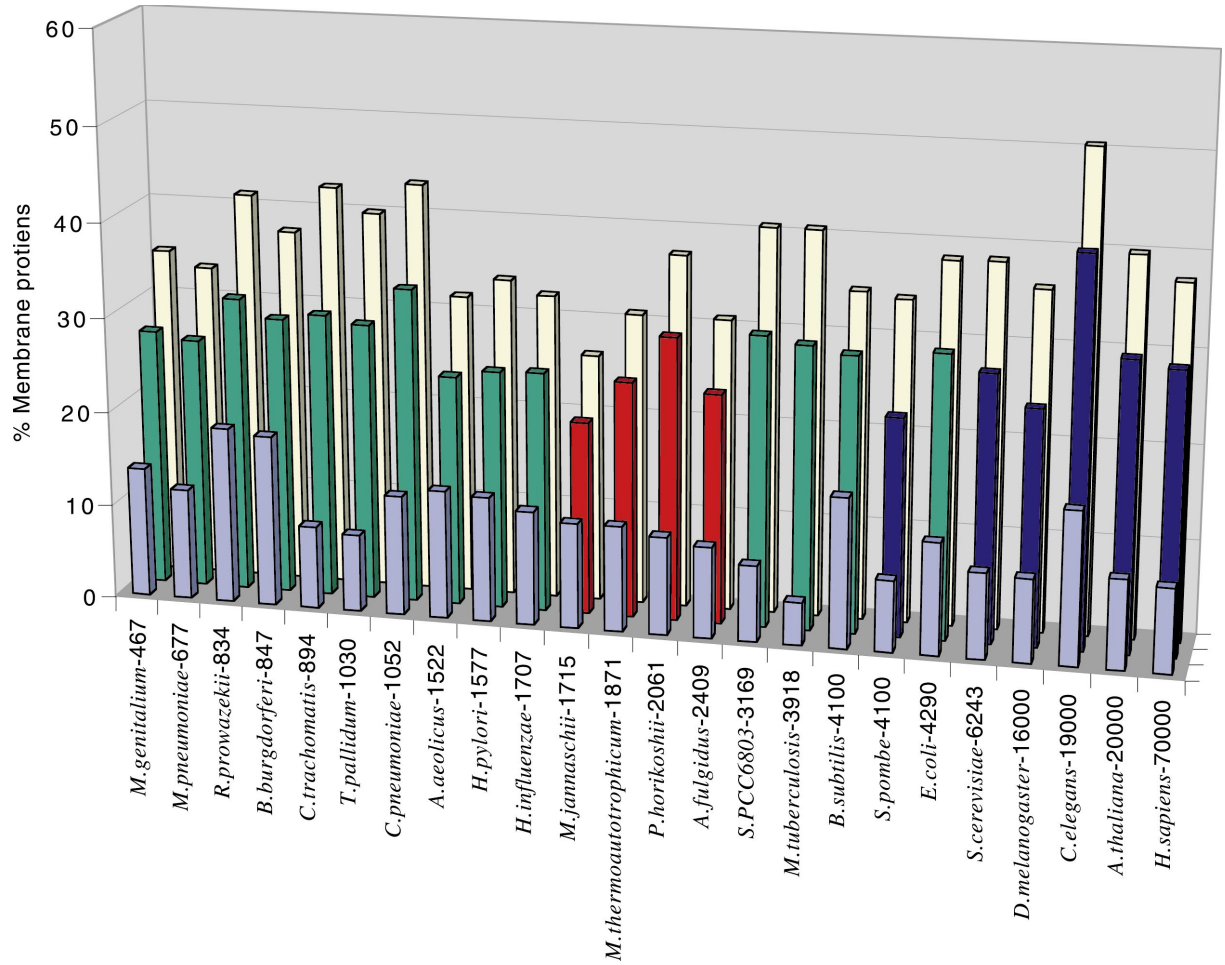

Fig. 2. A graph to show any relationship between the proportion of membrane proteins in a proteome and proteome size. The upper bound (yellow) represents the results of the permissive search. The lower hydropathy bound (purple), represents the restrictive search. The center histogram illustrates the optimized search results, where the results from members of archea, eubacteria, and eukaryota are colored red, green, and blue, respectively.

proteome size and the membrane protein fraction in any of the three contrasting searches one can be confident that any genome-specific search success is not clouding a general trend. Why *C. elegans* appears to have an unusual membrane protein complement is not obvious. However, one possibility is that the nematode worm, in

this aspect, is relatively unchanged from its early metazoan ancestor, which may have had diverse membrane function (considering its multicellularity), but used proportionately few, ancient, and perhaps ubiquitous aqueous pathways to support membrane processes. Although neural network–based algorithms are available for the the prediction of transmembrane α-helices from primary sequence data,[4] in this type of analysis they convey no significant benefits as compared to optimized hydropathy searching. Neural network prediction methods are based upon known transmembrane regions which are not particularly representative of any particular organism. Hence, these methods can be unreliable for a given organism and are notably worse for prokaryotes than eukaryotes.

CONCLUSION

Earlier studies had suggested that eukaryote genomes have a larger proportion of membrane proteins.[5] However, in these estimates the proportion of membrane proteins was compared to the size of the genome sequenced so far, not to the estimated size of the proteome or even the genome size. Also, we note that the idea of more complex organisms needing a disproportionately larger complement of membrane proteins is conceptually flawed. The basic idea about compartmentalization and cellular differentiation is that different chemistry takes place within each compartment or cell. Thus, whereas more membrane proteins are required to maintain intercompartmental communication, a complementary range of aqueous proteins is needed to undertake the particular biological function. Our data illustrate that overall, structurally and genomically more complex eukaryotes do not have any greater or lesser proportional requirement than eubacteria or archaebacteria for α-helical transmembrane proteins. In each of the three domains of life, membrane proteins are matched by an analogous complement of aqueous proteins.

REFERENCES

1. 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.
2. Sung CY, Gennity JM, Pollitt NS, Inouye M. A positive residue in the hydrophobic core of the *Escherichia coli* lipoprotein signal peptide suppresses the secretion defect caused by an acidic amino terminus. J Biol Chem 1992;267:997–1000.
3. Hikita C, Mizushima S. Effects of total hydrophobicity and length of the hydrophobic domain of a signal peptide on in vitro translocation efficiency. J Biol Chem 1992;267:4882–4888.
4. Rost B, Casadio R, Fariselli P, Sander C. Transmembrane helices predicted at 95-percent accuracy. Protein Sci 1995;4:521–533.
5. Wallin E, Von Heijne G. Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. Protein Sci 1998;7:1029–1038.