FEBS 19862 FEBS Letters 423 (1998) 281–285

Minireview

Do transmembrane protein superfolds exist?

David T. Jones*

Department of Biological Sciences, University of Warwick, Coventry CV4 7AL, UK

Received 7 January 1998

Abstract A reliable and widely used transmembrane protein structure prediction algorithm was applied to five representative genomic sequence data sets in order to re-examine the hypothesis that in contrast to globular proteins there are no favored transmembrane protein fold families. When the number of predicted membrane spanning segments and the topology of these segments is taken into account then definite biases are observed which suggest that certain transmembrane topologies are significantly more common than others.

Key words: Transmembrane protein; Protein structure prediction; Protein superfold

1. Introduction

Integral membrane proteins represent an important class of proteins which are employed in a wide range of cellular roles. The fact that these proteins are found in a lipid environment means that atomic resolution experimental structures for these proteins are few and far between. To date less than ten distinct integral membrane protein structures have been solved to atomic resolution, whereas several thousand structures for globular proteins are known. Fortunately, despite the difficulty in experimental structure determination for transmembrane proteins, the physicochemical constraints imposed by the lipid environment make the prediction of transmembrane protein structure somewhat more straightforward than for globular proteins. Over recent years a number of very reliable methods have been published for the prediction of transmembrane helices and also for the prediction of the chain topology with respect to the inside/outside membrane surfaces [1–4].

More recently, attention has been turning to the distribution of transmembrane proteins in the complete genomes for different organisms [4–6]. Although questions regarding the structure of proteins in different organisms can in general only be addressed by applying prediction methods, the reliability of modern transmembrane prediction methods is such that a reasonable degree of confidence in the resulting conclusions is possible. Here, a reliable transmembrane protein structure prediction method [2] was applied to five representative genomes in order to look for evidence that certain numbers of transmembrane spanning segments and topologies are highly favored.

2. Methods

To predict the structure and topology of transmembrane proteins, an expectation maximization method was used [2]. Briefly, the method classifies residues into five structural states as follows: L_i (inside loop), L_o (outside loop), H_i (inside helix end), H_m (helix middle), and H_o (outside helix end). The number of residues taken to be in the helix end caps was arbitrarily taken as being four. Using this definition of membrane protein topology, a set of statistical tables (log likelihood ratios, or log likelihoods for short) was compiled from well-characterized membrane protein data. The statistical tables show definite biases towards certain amino acid species on the inside, middle and outside of a cellular membrane. The most significant components of the propensities merely determine the lipophilic preferences of amino acids, or in other words that hydrophobic residues occur more frequently in the helical segments than the flanking regions. The signals that cannot be explained away by hydrophobicity alone are perhaps of more interest. The previously described preference for positively charged residues to be found in the inside loops is clearly seen, but it is also interesting to note that a similar effect is seen between the inside and outside helix caps, though this could be due to errors in assigning the boundaries between the membrane spanning segments and their flanking regions. For multi spanning proteins, the most significant preferences for inside/outside loop are Arg, Gly, His, Lys and Pro, whereas for single spanning proteins, Ala, Arg, Asp, Gln, Lys, Pro, Thr, Trp and Val have the most significant propensities. For inside/outside helix caps, only Phe and Trp have highly significant topogenic propensities for multi spanning helices whereas Cys, Gly, His, Leu, Lys, Phe, Pro, Ser, Thr, Tyr and Val show clear inside/outside preferences for single spanning helix caps.

To determine the most likely transmembrane model for each target sequence, a dynamic programming method was used, similar in principle to dynamic programming sequence alignment algorithms. This algorithm was used to find the best set of variables (number, position, length and direction) for each considered model.

To identify probable signal peptides, a log likelihood sequence profile spanning 20 amino acid positions was calculated based on the most reliable signal peptide annotations for secreted proteins in SWISSPROT Release 34. This profile was scanned across the first 50 residues of each gene product and the highest scoring window position calculated. Where the maximum window score exceeded a score of 7.5 the 20 residue peptide was masked out from the calculation of topological models.

*Fax: +44 (1203) 523568.
E-mail: jones@globin.bio.warwick.ac.uk

PII S0014-5793(98)00095-7

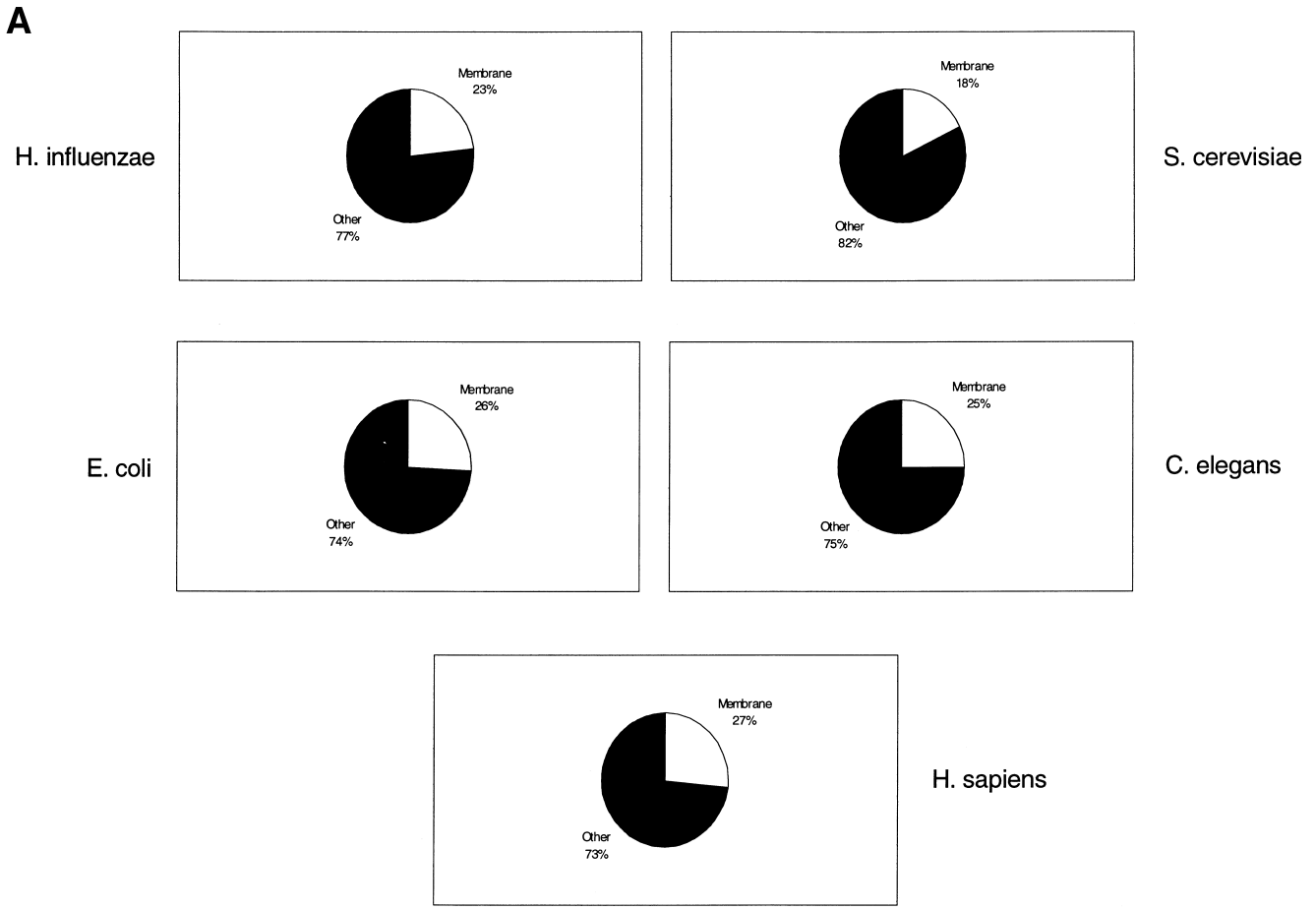


Fig. 1. A: Fraction of open reading frames with predicted transmembrane spanning segments taken from five representative genomes. Total number of ORFs analyzed in each genome were as follows: 1696 (*H. influenze*), 4290 (*E. coli*), 6181 (*S. cerevisiae*), 7299 (*C. elegans*), 4056 (*H. sapiens*). B: Distribution of predicted transmembrane topologies for the five genomes.

3. Proportion of transmembrane proteins in representative genomes

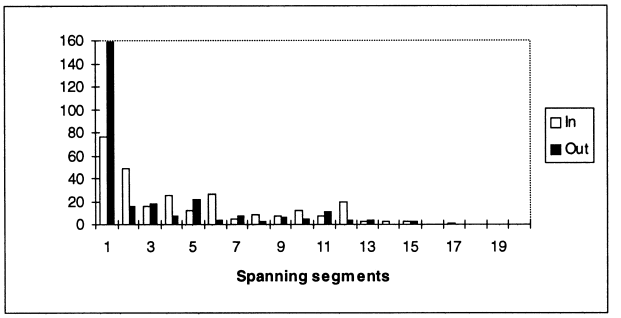
The first question that comes to mind is regarding the proportion of the gene products for each genome which code for transmembrane proteins. Averaged over all the available genomes, there is some variance in the estimates, where values from 20 to 40% have been reported. Simple hydropathy analysis of protein sequences tends to overpredict membrane spanning segments [2], and so it is not surprising that the higher estimates are based on these simpler methods. Fig. 1A shows the proportions estimated using an enhanced version of MEMSAT [2] on five representative genomes. In this instance the estimates (19% averaged over all five genomes) are very much at the lower end of the range of previous estimates. This is partly due to the use of an accurate prediction method, but is mostly due to additional consideration of signal peptides in this analysis. Cleavable signal peptides which enable the secretion of certain polypeptides into their required cellular compartment are frequently mistaken for transmembrane segments [2], and in this case a simple signal peptide prediction method was applied as a preprocessing step. Of course, like other estimates, the above estimates do not account for the possible occurrence of predominantly beta-sheet transmembrane proteins such as porin, but based on spectroscopic analyses of known integral membrane proteins the beta-sheet class of transmembrane proteins is probably rare [7]. Nonetheless, it must be expected that a certain number of predicted gene products from the currently completed genomes will have not been correctly predicted as integral membrane proteins in currently published annotations. Despite these reservations it does appear from Fig. 1A that the proportion of transmembrane proteins in genomes is relatively constant across quite different organisms, but given the small fraction of the human genome being considered here it is really too early to come to any firm conclusions. Clearly the most complex organisms might well have a greater need for transmembrane proteins to fulfil roles in inter-cellular signalling or immune response, but there is little evidence for this at present.

4. Distribution of predicted transmembrane topologies

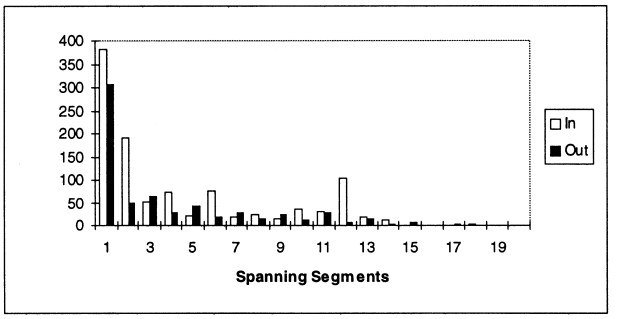
Two recent papers have looked more closely at the predicted transmembrane proteins [5,6], and have classified the predicted proteins by the number of predicted membrane spanning segments. Proteins with single membrane spanning segments should properly be considered as a separate class from those with multiple spanning segments [8], as these are, very generally speaking, globular protein domains which are simply anchored to the membrane by a hydrophobic helix. Indeed these membrane anchors are often referred to as uncleaved signal peptides. Both these analyses show a similar

B

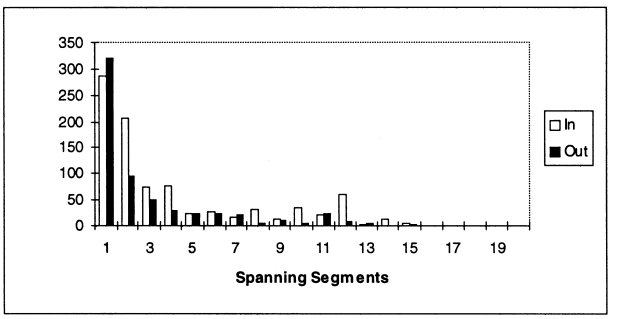
H. influenzae



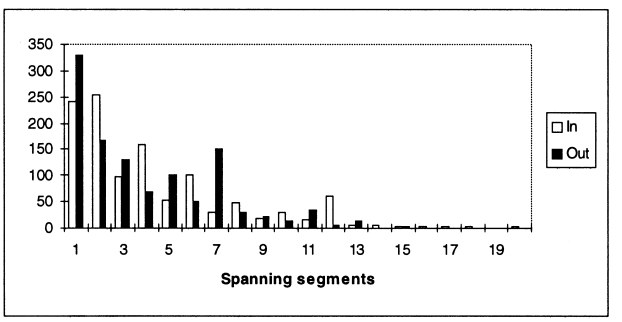
E. coli



S. cerevisiae



C. elegans



H. sapiens

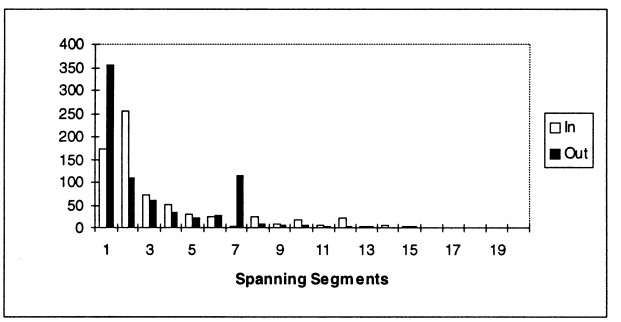


Fig. 1 (*continued*)

trend where the fraction of structures decreases monotonically from the most populated single spanning segment category to the rare families with many spanning segments. This trend has been interpreted as indicating that there are no predominant families of membrane protein folds. In globular proteins, certain folds (e.g. the TIM barrel or the immunoglobulin fold) are seen to recur frequently in unrelated families of proteins. These folds have been termed 'superfolds' [9], and ten such folds have been defined so far [10]. That membrane protein structures should not have equivalents of superfolds is somewhat surprising, but nonetheless there does appear to be little evidence from the previous published analyses for predominance of particular numbers of transmembrane segments.

Fig. 1B shows a somewhat different view of the distribution of predicted transmembrane protein structures in the same five representative genomes. The difference between this analysis and those mentioned above is that here the topology of the structures has been predicted in addition to simply counting the number of predicted spanning segments. This additional level of classification changes the picture somewhat. When the topology of the structures is taken into account, a small number of predominant topological families do present themselves. The decreasing trend which favors small numbers of spanning segments is still apparent, but this is really just a reflection of the distribution of protein chain lengths. Above this background distribution can be seen some unexpectedly populated classes. Using a simple shorthand notation where, for example, 5in represents a structure with five spanning segments, with the N-terminus directed towards the inside (cytoplasm) of the cell, the bacterial transmembrane 'supertopologies' are apparently: 4in, 6in and 12in, with strong bias also seen for 5out, 10in and 14in. For yeast, favored topologies are: 4in, 8in, 10in, 12in and 14in. Taking microbial genomes as a group, there appear to be three highly populated topology classes: 4in, 10in and 12in, with particular bias being shown for 4in and 12in. Interestingly, a different pattern is observed for multicellular organisms. For *C. elegans* and a small sample of the human genome there appear to be two favored topologies: 4in and 7out. *C. elegans* also shows again a strong bias for the ubiquitous 12in topology. Given the small fraction of the human genome being considered here, it is too early to say conclusively that it does not also show a bias for 12in.

5. Evolutionary arguments for the observed distribution of transmembrane topologies

Given the lack of experimental data for transmembrane protein structures, it is only possible to speculate as to whether distant evolutionary relationships are responsible for the observed bias in the distribution of predicted transmembrane protein topologies. However, a number of prominent transmembrane protein families can be found in the highly populated topology classes. The most obvious superfamily is of course that of the seven transmembrane helix G-coupled receptors, which may or may not share a distant relationship with the bacteriorhodopsin superfamily. In both cases the transmembrane topology is known to be 7out, but the packing arrangement of the helices is thought to differ between these superfamilies [11]. Very prominent members of the 12in topological family are of course the various permease proteins, including amongst others the lactose-proton symport protein. Another very significant superfamily which also falls into the 12in topological class includes the various ABC transporter proteins, which is widespread in both prokaryotes and eukaryotes, and in humans includes the cystic fibrosis conductance regulator protein. The universal bias towards the 4in topology is somewhat harder to explain. Interestingly there were very few 4in topologies in the set of proteins used in the original testing of the prediction method used here, and the only examples were the gap proteins from human and frog, and the colicin A immunity protein from *E. coli*, and no obvious relationships could be observed between the protein family names which comprised the 4in topological class. As with globular protein superfolds, it is not unthinkable that some of the biases observed in the distribution of transmembrane topologies are a result not of distant evolutionary relationships but merely stem from the physical constraints of protein folding. Though most probably coincidental, it is nonetheless interesting to note that the 4-helix bundle is also a globular protein superfold.

6. Random sequence tests

One concern that comes to mind in making these observations is that the observations are simply due to artefacts in the prediction method used. This is of course a possibility, but as a simple control experiment, the sequences from *Haemophilus influenzae* were shuffled in blocks so as to preserve the sequential patterns of hydrophobicity and these shuffled sequences were then fed into the topology prediction program again. Although the resulting predictions for the shuffled sequences showed the expected distribution in terms of number of predicted transmembrane segments (correlating with sequence length), no bias was apparent between inside and outside topologies. This suggests that the observed topological biases are not an obvious result of intrinsic biases in the prediction method used.

7. Conclusions

The results here suggest that whilst the proportion of transmembrane proteins is relatively constant across five quite different organisms, there are nonetheless some biases in the distribution of transmembrane topologies which might suggest the existence of transmembrane 'superfolds', or at least highly populated fold families. In addition, these biases are not readily explained simply from the distributions of sequence lengths in the genomes. It must be stressed, however, that these observations of biases in the distributions of predicted transmembrane topology represent a very superficial view of the real situation. Without many more experimentally determined 3-D structures for transmembrane proteins it is impossible to be confident about the existence of highly populated fold families in these proteins. Nevertheless these observations suggest that in the absence of a significant quantity of experimental structural information, there may be something to gain from attempts to classify integral membrane proteins into broad structural families, perhaps based on prediction methods (as used here) or conserved sequence motifs. It may also be the case that by trying to rationalize the topological biases, improvements may also be made to the prediction methods themselves.

References

[1] Sipos, L. and von Heijne, G. (1993) Eur. J. Biochem. 213, 1333–1340.
[2] Jones, D.T., Taylor, W.R. and Thornton, J.M. (1994) Biochemistry 33, 3038–3049.
[3] Persson, B. and Argos, P. (1996) Protein Sci. 5, 363–371.
[4] Rost, B., Fariselli, P. and Casadio, R. (1996) Protein Sci. 5, 1704–1718.
[5] Frishman, D. and Mewes, H.W. (1997) Nature Struct. Biol. 4, 626–628.
[6] Arkin, I.T., Brunger, A.T. and Engelman, D.M. (1997) Proteins 28, 465–466.
[7] Deber, C.M. and Li, S.C. (1995) Biopolymers 37, 295–318.
[8] Whitley, P., Saaf, A., Gafvelin, G., Johansson, M., Wallin, E. and von Heijne, G. (1995) Biochem. Soc. Trans. 23, 965–967.
[9] Orengo, C.A., Jones, D.T. and Thornton, J.M. (1994) Nature 372, 631–634.
[10] Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B. and Thornton, J.M. (1997) Structure 5, 1093–1108.
[11] Baldwin, J.M., Schertler, G.F.X. and Unger, V.M. (1997) J. Mol. Biol. 272, 144–164.