ELSEVIER

Biophysical Chemistry 82 (1999) 165–171

Biophysical Chemistry

www.elsevier.nl/locate/bpc

Proportion of membrane proteins in proteomes of 15 single-cell organisms analyzed by the SOSUI prediction system

Shigeki Mitaku*, Mitsuo Ono, Takatsugu Hirokawa[1], Seah Boon-Chieng, Masashi Sonoyama

Tokyo University of Agriculture and Technology, Department of Biotechnology, Koganei, Tokyo 184-8588, Japan

Received 15 September 1999; accepted 15 September 1999

Abstract

A software system, SOSUI, was previously developed for discriminating between soluble and membrane proteins and predicting transmembrane regions (Hirokawa et al., Bioinformatics, 14 (1998) 378–379). The performance of the system was 99% for the discrimination between two types of proteins and 96% for the prediction of transmembrane helices. When all of the amino acid sequences from 15 single-cell organisms were analyzed by SOSUI, the proportion of predicted polytopic membrane proteins showed an almost constant value of 15–20%, irrespective of the total genome size. However, single-cell organisms appeared to be categorized in terms of the preference of the number of transmembrane segments: species with small genomes were characterized by a significant peak at a helix number of approximately six or seven; species with large genomes showed a peak at 10 or 11 helices; and species with intermediate genome sizes showed a monotonous decrease of the population of membrane proteins against the number of transmembrane helices. © 1999 Elsevier Science B.V. All rights reserved.

Keywords: Genome; Proteome; Membrane protein; Bioinformatics; Prediction

* Corresponding author.

[1] Present address. Ryoka Inc., Computational Science and Technology Division, Irifune, Urayasu, Chiba Pref. 279-0012, Japan.

PII: S0301-4622(99)00116-7

1. Introduction

The targets of many medicinal reagents are intrinsic membrane proteins, such as receptors and transporters. However, it is very difficult to identify experimentally a target membrane protein, which binds a certain ligand, from the complete set of amino acid sequences in a proteome, because the number of open reading frames (ORFs) is very large. The number of ORFs that have to be examined for the identification of the target membrane protein would be greatly reduced, however, if the amino acid sequences were accurately classified into soluble and membrane proteins by some computational method. Furthermore, if the number of transmembrane helices can be determined accurately, then the number of possible candidates of target membrane proteins may be reduced further, using the correlation between the helix numbers of membrane proteins and their functions [1,2]. Therefore, an accurate method to discriminate between soluble and membrane proteins and to predict transmembrane segments is one of the most important theoretical approaches for analyses of the amino acid sequences from total genomes [3–6].

The discrimination of membrane proteins from other types of proteins is closely related to the prediction of transmembrane segments. Most computational methods to predict transmembrane helices are based upon the fact that transmembrane helices have high hydrophobicity [7,8]. The accuracy of prediction has been improved by various supplementary parameters [9–17]. However, those methods were not intended to distinguish primarily between soluble and membrane proteins and their accuracy did not exceed that of the method by Klein et al. [18], which attained an accuracy of 95%. This method is based on the fact that most membrane proteins have at least one very hydrophobic segment, which effectively allows their discrimination from the hydrophobic segments in soluble proteins.

In spite of the superior performance of the previous method, a consideration of the misprediction of the amino acid sequences of soluble proteins in a proteome highlights the need for further improvement of the accuracy. Membrane proteins are a minor component of the proteome. Therefore, the false positive prediction of transmembrane segments in soluble proteins amplifies the misprediction (overprediction) of membrane proteins. If membrane proteins are only one-fifth of the total ORFs, then a 5% misprediction 5% in the analysis of soluble proteins will lead to an overprediction of membrane proteins of as large as 20%.

We recently developed a physicochemical method for membrane protein prediction, the SOSUI system, with accuracy of 99% for the discrimination between soluble and membrane proteins and 96% for the prediction of transmembrane helices [19]. Here, we describe analyses of the amino acid sequences from the total genomes of 15 single-cell organisms, including 11 eubacteria, three archaea, and one eukarya. The results

Table 1
List of single-cell organisms with the number of open reading frames

Eubacteria	Number of ORFs	Archaea	Number of ORFs
Mycoplasma genitalium	467	*Methanococcus jannaschii*	1715
Mycoplasma pneumoniae	677	*M. thermoautotrophicm*	1871
Chlamydia trachomatis	894	*Archaeoglobus fulgidus*	2049
Treponema pallidum	1031		
Borrelia burgdorferi	1638	Eukarya	
Aquifex aeolicus	1522	*Saccharomyces cerevisiae*	6217
Helicobacter pylori	1577		
Haemophilus influenzae	*1713*		
Synechocystis sp.	3169		
Bacillus subtilis	4099		
Escherichia coli	4290		

indicate that the proportion of membrane proteins in single-cell organisms is almost constant, in the range from 15 to 20%, and that there were three types of species in terms of the preference of the number of transmembrane helices.

2. Materials and methods

The open reading frames (ORFs) of 15 species were analyzed in this work (Table 1). The eubacteria are: *Mycoplasma genitalium* [20]; *Mycoplasma pneumoniae* [21]; *Chlamydia trachomatis* [22]; *Treponema pallidum* [23]; *Borrelia burgdorferi* [24]; *Aquifex aeolicus* [25]; *Helicobacter pylori* [26]; *Haemophilus influenzae* [27]; *Synechocystis* sp. [28]; *Bacillus subtilis* [29]; and *Escherichia coli* [30]. The archaea are: *Methanococcus jannaschii* [31]; *Methanobacterium thermoautotrophicm* [32]; and *Archaeoglobus fulgidus* [33], and the eukaryotic single-cell organism is *Saccharomyces cerevisiae* [34]. All ORFs extracted from the genome sequences were downloaded from the appropriate WWW sites.

All of the amino acid sequences of the 15 species were analyzed by the software system SOSUI, which classifies sequences into two types of proteins, soluble and membrane, and predicts the transmembrane segments in membrane proteins. This system uses three physicochemical parameters for the analysis: the hydrophobicity of the segments, the length of the proteins, and the preference of detergent-like residues at the ends of transmembrane helical segments [19]. The detergent-like residues are Lys, Arg, His, Glu, Gln, Trp, and Tyr, which have a polar group as well as long stem chain. The accuracy of the discrimination between soluble and membrane proteins is as good as 99%, and that of the transmembrane helix prediction is approximately 96% [19].

3. Results and discussion

3.1. Fraction of membrane proteins in all ORFs

The single-cell organisms studied in this work have a wide range of genome sizes from 0.58 (*Mycoplasma genitalium*) to 12.0 Mb (*Saccharomyces cerevisiae*). The corresponding range of the number of ORFs was between 467 and 6217. Fig. 1 shows the number of membrane proteins predicted by SOSUI as a function of the total number of ORFs. The species are very dif-

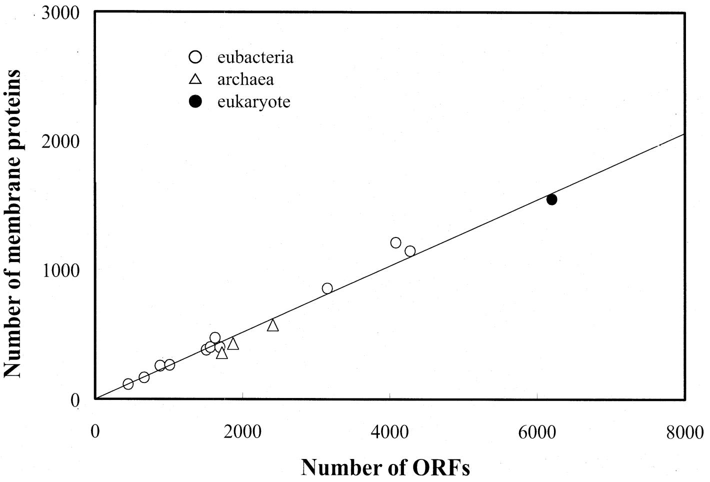

Fig. 1. The number of membrane proteins, predicted by the SOSUI system, with at least two transmembrane segments as a function of the total number of all ORFs in 15 proteomes. The 15 single-cell organisms include: 11 eubacteria (open circles); three archaea (open triangles); and one eukaryote (solid circle). The solid line represents a proportion of membrane proteins of approximately 17%.

ferent, from the evolutionary point of view, and include 11 eubacteria, three archaea, and one eukarya. It was found that the scattering of data from a linear relationship was very small. Namely, the fraction of membrane proteins was almost constant, in spite of the large differences in the genome sizes. The fraction of membrane proteins was between approximately 15 and 20%, whereas, the fraction of total membrane proteins was approximately 25%. We have plotted the number of predicted membrane proteins with at least two transmembrane helices in Fig. 1, which shows the proportion of polytopic membrane proteins. The error in the prediction of polytopic membrane proteins was estimated to be smaller than 2%, from the product of the fraction of single spanning membrane proteins (< 30%) and the error of transmembrane helix prediction (< 5%) in our system.

There is some discrepancy between the present result and previous predictions. Frishman and Mewes [3] reported the fraction to be approximately 35% from analyses of the total genomes of five bacteria. Their value is considerably larger than the present result. Wallin and von Heijne [5,35] comprehensively studied total ORFs and

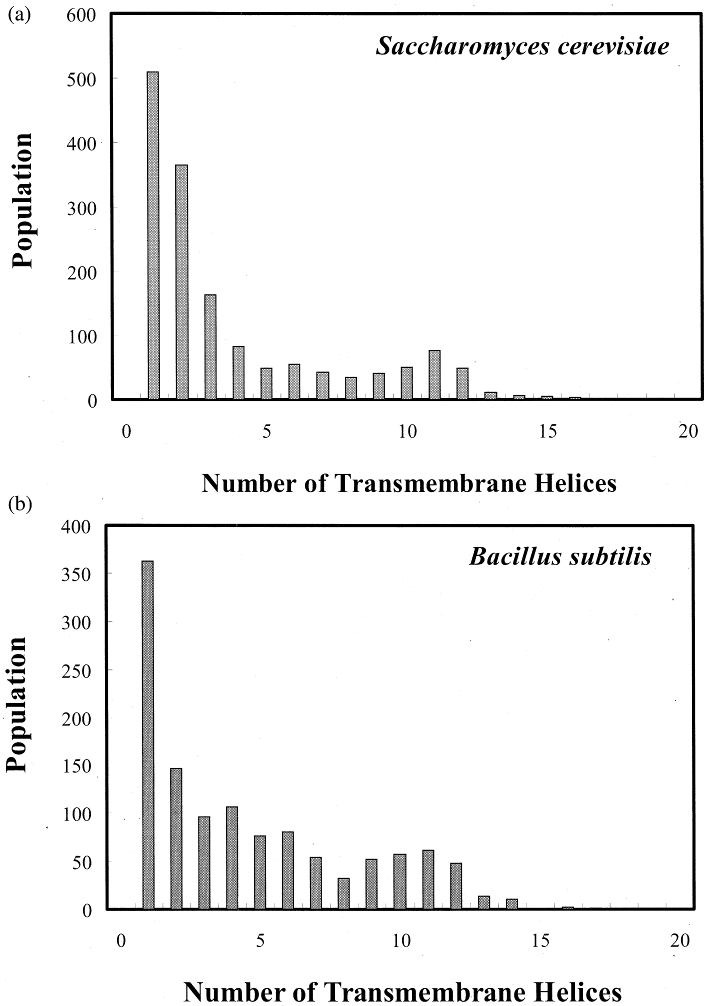

Fig. 2. Fraction of membrane proteins with different numbers of predicted transmembrane segments. The histograms of the transmembrane helix numbers are shown for four species with various genome sizes: (a) *Saccharomyces cerevisiae* (type 3); (b) *Bacillus subtilis* (type 3); (c) *Methanococcus jannaschii* (type 2); and (d) *Mycoplasma pneumoniae* (type 1).

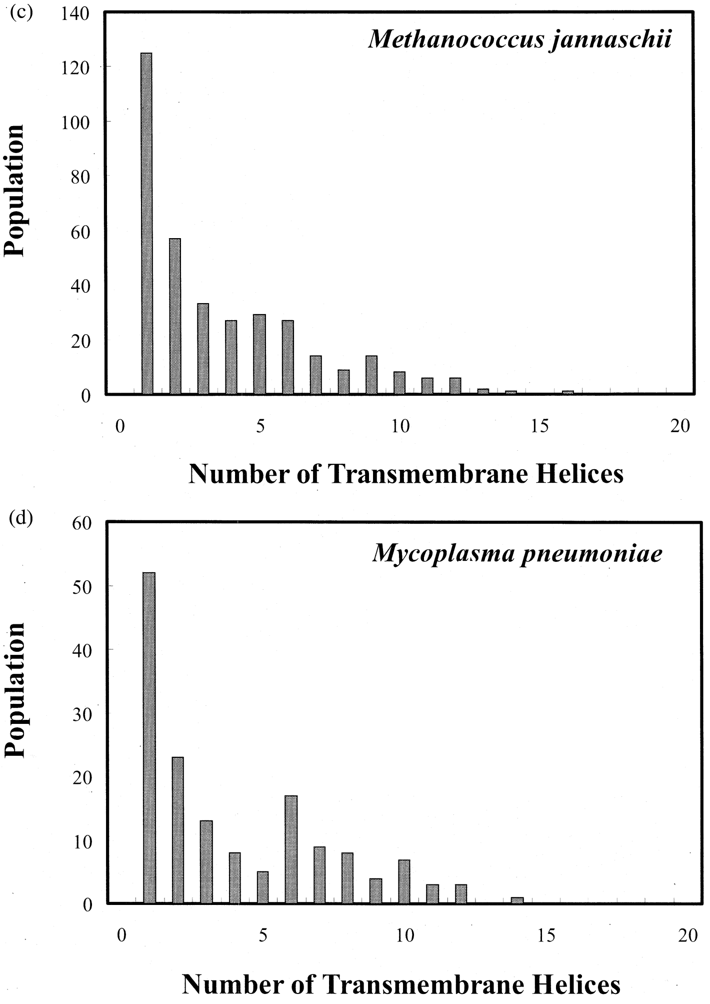

Fig. 2. (*Continued*)

discriminated membrane proteins with more than two transmembrane segments. Of all the ORFs, 20–30% were predicted to encode such membrane proteins. This number of membrane proteins is also larger than the value in this work. Jones [6] reported a smaller fraction of membrane proteins, after analyzing a small dataset of five proteomes. In this work, we have analyzed 15 proteomes of single cell organisms and obtained a fraction of 15–20%, which is similar to the results of Jones. However, the constant proportion of membrane proteins does not necessarily mean that the fraction of membrane proteins with the same function (e.g. receptor or channel) is constant among the 15 different species, as described in the next section.

The analysis by Wallin and von Heijne [5] indicated a higher fraction in the larger genomes than in the smaller ones, in contrast to the constant fraction of membrane proteins in this work. Unless there is some systematic error in the prediction of membrane proteins, this sort of discrepancy may not be understood. However, the dependence of prediction error on the genome size remains to be studied in the future.

3.2. Number of transmembrane segments in membrane proteins

Transmembrane segments were predicted for all membrane proteins in the fifteen proteomes by the SOSUI system. Fig. 2a–d show the population of membrane proteins with different numbers of transmembrane helices for *Saccharomyces cerevisiae* (6217 ORFs), *Bacillus subtilis* (4099 ORFs), *Methanococcus jannaschii* (1715 ORFs), and *Mycoplasma pneumoniae* (677 ORFs), respectively. The histogram of the transmembrane helix number for *S. cerevisiae* has a significant peak at 11 helices (Fig. 2a), which agrees with the results of Wallin and von Heijne [5]. *Bacillus subtilis* (Fig. 2b), as well as *Escherichia coli*, also showed a peak at the helix number of 11. However, few significant peaks were observed in the helix number histograms for bacteria with 1000–2000 ORFs. For example, *Methanococcus jannaschii* showed a very small maxima at helix numbers of five and nine (Fig. 2c). *M. pneumoniae*, which has only 677 ORFs, exhibited a significant peak at the helix number of six (Fig. 2d) Thus, single-cell organisms could be classified into three types of helix number distributions. Species with small genome sizes seem to have a peak at six or seven transmembrane helices (type 1). A monotonous decrease against the helix number is observed for species with intermediate genome sizes (type 2). Single cell organisms with large genomes, including *S. cerevisiae*, have a peak at 10–12 transmembrane helices (type 3).

However, it is difficult to discuss the functional aspects of these characteristic distributions of the transmembrane helix numbers. It is well known that G-protein-coupled receptors have seven transmembrane helices [1], and that many membrane proteins with more than 10 transmembrane helices are transporters [2]. Therefore, if we can precisely predict the number of transmembrane helices in membrane proteins, then we will be able to discuss the constitution of membrane proteins in a proteome from the viewpoint of their functions. However, the accuracy of transmembrane helix prediction in this work was approximately 96%, and the possibility of predicting all 10 transmembrane helices correctly decreases to only two-thirds ($0.96^{10} = 0.66$). A predicted membrane protein with six transmembrane helices may actually have seven transmembrane segments, and vice versa, as determined by the present systems. Therefore, we have to improve the accuracy of the transmembrane helix prediction to infer the function of a membrane protein from the number of transmembrane helices.

References

[1] S. Watson, S. Arkinstall, The G-Protein Linked Receptor Facts Book, The Academic Press, London, 1996.

[2] J. Griffith, C. Sanson, The Transporter Facts Book, The Academic Press, London, 1997.

[3] D. Frishman, H.W. Mewes, Protein structural classes in five complete genomes, Nat. Struct. Biol. 4 (1997) 625–628.

[4] I.T. Arkin, A.T. Brunger, D.M. Engelman, Are there dominant membrane protein families with a given number of helices? Proteins 28 (1997) 465–466.

[5] E. Wallin, G. von Heijne, Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms, Protein Sci. 7 (1998) 1029–1038.

[6] D.T. Jones, Do transmembrane protein superfolds exist? FEBS Lett. 423 (1998) 281–285.

[7] T.A. Steitz, A. Goldman, D.M. Engelman, Quantitative application of the helical hairpin hypothesis to membrane proteins, Biophys. J. 37 (1982) 124–125.

[8] J. Kyte, R.F. Doolittle, A simple method for displaying the hydropathic character of a protein, J. Mol. Biol. 157 (1982) 105–132.

[9] D. Eisenberg, R.M. Weiss, T.C. Terwilliger, The hydrophobic moment detects periodicity in protein hydrophobicity, Proc. Natl. Acad. Sci. USA 81 (1984) 140–144.

[10] S. Mitaku, S. Hoshi, T. Abe, R. Kataoka, Spectral analysis of amino acid sequence. I. Intrinsic membrane proteins, J. Phys. Soc. Jpn. 53 (1984) 4083–4090.

[11] S. Mitaku, S. Hoshi, R. Kataoka, Spectral analysis of amino acid sequence. II. Characterization of ?-helices by local periodicity, J. Phys. Soc. Jpn. 54 (1985) 2047–2054.

[12] D.C. Rees, L. DeAntonio, D. Eisenberg, Hydrophobic organization of membrane proteins, Science 245 (1989) 510–513.

[13] F. Jähnig, Structure prediction of membrane proteins are not that bad, TIBS 15 (1990) 93–95.

[14] D.T. Jones, W.R. Taylor, J.M. Thornton, A model recognition approach to the prediction of all-helical membrane protein structure and topology, Biochemistry 33 (1994) 3038–3049.

[15] G. von Heijne, Membrane protein structure prediction

— hydrophobicity analysis and the positive-inside rule, J. Mol. Biol. 225 (1992) 487–494.

[16] B. Persson, P. Argos, Prediction of transmembrane segments in proteins utilizing multiple sequence alignments, J. Mol. Biol. 237 (1994) 182–192.

[17] E. Hartmann, T.A. Rapoport, H.F. Lodish, Predicting the orientation of eukaryotic membrane-spanning proteins, Proc. Natl. Acad. Sci. USA 86 (1989) 5786–5790.

[18] P. Klein, M. Kanehisa, C. DeLisi, The detection and classification of membrane-spanning proteins, Biochim. Biophys. Acta 815 (1985) 468–476.

[19] T. Hirokawa, S. Boon-Chieng, S. Mitaku, SOSUI: classification and secondary structure prediction system for membrane proteins, Bioinformatics 14 (1998) 378–379.

[20] C.M. Fraser, J.D. Gocayne, O. White et al., The minimal gene complement of *Mycoplasma genitalium*, Science 270 (1995) 397–403.

[21] R. Himmelreich, H. Hilbert, H. Plagens, E. Pirkl, B.C. Li, R. Herrmann, Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*, Nucleic Acids Res. 24 (1996) 4420–4429.

[22] R.S. Stephens, S. Kalman, C. Lammel et al., Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*, Science 282 (1998) 754–759.

[23] C.M. Fraser, S.J. Norris, G.M. Weinstock et al., Complete genome sequence of *Treponema pallidum*, the syphilis spirochete, Science 281 (1998) 375–388.

[24] C.M. Fraser, S. Casjens, W.M. Huang et al., Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*, Nature 390 (1997) 580–586.

[25] G. Deckert, P.V. Warren, T. Gaasterland et al., The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*, Nature 392 (1998) 353–358.

[26] J.F. Tomb, O. White, A.R. Kerlavage et al., The complete genome sequence of the gastric pathogen *Helicobacter pylori*, Nature 388 (1997) 539–547.

[27] R.D. Fleishmann, M.D. Adams, O. White et al., Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd, Science 269 (1995) 496–512.

[28] T. Kaneko, S. Sato, H. Kotani et al., Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions, DNA Res. 3 (1996) 109–136.

[29] F. Kunst, N. Ogasawara, I. Moszer et al., The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*, Nature 390 (1997) 249–256.

[30] F.R. Blattner, G. Plunkett III, C.A. Bloch et al., The complete genome sequence of *Escherichia coli* K-12, Science 277 (1997) 1453–1474.

[31] C.J. Bult, O. White, G.J. Olsen et al., Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*, Science 273 (1996) 1058–1073.

[32] D.R. Smith, L.A. Doucette-Stamm, C. Deloughery et al., Complete genome sequence of *Methanobacterium thermoautotrophicum* deltaH: functional analysis and comparative genomics, J. Bacteriol. 179 (1997) 7135–7155.

[33] H.P. Klenk, R.A. Clayton, J.F. Tomb et al., The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*, Nature 390 (1997) 364–370.

[34] A. Goffeau et al., Nature 387 (Suppl.) (1997) 5.

[35] M.G. Claros, G. von Heijne, TopPred II: an improved software for membrane protein structure prediction, CABIOS 10 (1994) 685–686.