Detection of Membrane Proteins in the Whole Genome Sequences

Daisuke Kihara kihara@kuicr.kyoto-u.ac.jp
Minoru Kanehisa kanehisa@kuicr.kyoto-u.ac.jp
Institute for Chemical Research, Kyoto University
Gokasho, Uji, Kyoto 611, Japan

Abstract

Membrane proteins are detected from whole genome sequences of the nine organisms which were determined recently. Transmembrane segments are detected using the prediction method reported previously by us and the distribution of groups of proteins sharing a specific number of transmembrane segments are compared among the organisms.

1 Introduction

As the genome sequencing projects progress rapidly, complete genome sequences of several organisms are now available. It is an indispensable process to assign the biological meaning (function, structure) to the sequences and to analyze arrangements and compositions of genes in the whole genomes. This new type of sequence analysis which handles whole genome sequences is expected to reveal the structures of the genomes, mechanisms of gene expression, and detailed evolutional relationships among the organisms.

In the present study we have focused on membrane proteins. Membrane proteins are paid special attention in the genome analysis [1, 2], because they are involved in membrane transport and energy synthesis, so there is the possibility that their arrangement in the genomes directly reflects the biosynthetic potential of the organisms and the environment in which they inhabit. We used nine complete genome sequences, those of *Methanococcus jannaschii* and *Methanobacterium thermoautotrophicum* from Archaea, *Escherichia coli*, *Haemophilus influenzae*, *Helicobacter pylori*, *Mycoplasma genitalium*, *Mycoplasma pneumoniae* and *Synechocystis* sp. from Bacteria, *Saccharomyces cerevisiae* from Eukarya. ORF sequences of membrane proteins are detected mainly using discriminant analysis and transmembrane segments are detected in those sequences. The amount of membrane proteins in the genomes, distribution of groups of proteins sharing a specific number of transmembrane segments are analyzed and compared among the organisms.

2 Method

2.1 Data

The nine genome sequences are taken from KEGG [4]. The training set of membrane protein sequences were extracted from SWISS-PROT *rel.* 34.0. Fragment entries are excluded. One sequence from each 30% identity family was used and the set consisted of 3251 sequences. Another training set, the set of globular protein sequences are based on the PDBSELECT database 97-March version [3]. 35% threshold list was used excluding entries of membrane or lipid associated proteins (1AXN, 1COLA, 1HGEA, 1OCC, 1PRC, 1IDO, 2POR, 1ATY, 1SPF). The set consisted of 928 sequences.

2.2 Detection of ORF Sequences of Membrane Proteins

We mainly employed discriminant analysis to detect membrane protein sequences. The discrimination function was constructed to discriminate the most hydrophobic 21 residue-long regions of sequences from two groups of the training set, membrane proteins and globular proteins. The formula is as following: $f = -10.62 + 6.91 < H >$, where $< H >$ is the average hydrophobicity of regions using the index proposed by Kyte & Doolittle [6]. ORF sequences were considered to be of membrane proteins if the function was positive. 93.54% of membrane proteins and 95.58% of globular proteins in the training sets were assigned correctly. As the next step, possibilities of signal sequences and homology to membrane protein sequences in databases were also taken into consideration to reduce false assignments. Once a sequence was determined to be of a membrane protein, the prediction method for detecting transmembrane segments presented by us previously [5] was applied.

3 Results and Discussions

The rough estimation of the amount of membrane proteins in the genomes of the nine organisms applying the discrimination function is as follows: *Methanococcus jannaschii*: 23.51%, *Methanobacterium thermoautotrophicum*: 26.99%, *Escherichia coli*: 34.57%, *Haemophilus influenzae*: 26.67%, *Helicobacter pylori*: 28.12%, *Mycoplasma genitalium*: 30.19%, *Mycoplasma pneumoniae*: 28.95%, *Synechocystis* sp.: 33.35%, *Saccharomyces cerevisiae*: 30.29%. The actual number of membrane proteins may be smaller because there are some proteins with highly hydrophobic signal sequences. Though the refinement of the estimation is needed, it is probable that Archaea have relatively small amount of membrane proteins. Further analysis of the genome scale arrangements of membrane proteins will reveal the difference between Archaea and other organisms. We are now in the process of assigning functions to those membrane proteins using homology and motif search.

Acknowledgments

The authors thank Takatsugu Hirokawa for useful suggestions. This work was supported in part by a Grant-in-Aid for Scientific Research on Priority Areas, 'Genome Science', from the Ministry of Education, Science, Sports and Culture in Japan. The computation time was provided by the Supercomputer Laboratory, Institute for Chemical Research, Kyoto University.

References

[1] Arkin, I.T., Brünger, A.T, Engelman, D.M., "Are there dominant membrane protein families with a given number of helices?, " *Proteins*, 28:465–466, 1997.

[2] Clayton, R.A., White, O., Ketchum, A., Venter, J.C., "The first genome from the third domain of life," *Science*, 387:459–462, 1997.

[3] Hobohm, U., Scharf, M., Schneider, R., Sander, C., "Selection of a representative set of structures from the Brookhaven Protein Data Bank, " *Prot. Sci.*, 1:409–417, 1992.

[4] Kanehisa, M., "A database for post-genome analysis," *Trends. Genet.*, 13:375-376, 1997.

[5] Kihara, D., Shimizu, T., Kanehisa, M., "A prediction method for transmembrane segments in proteins utilizing multiple discrimination functions, " *Genome Informatics 1996*, pp.244-245, Universal Academy Press, Tokyo, 1996.

[6] Kyte, J., Doolittle, R.F., "A simple method for displaying the hydropathic character of a protein, " *J. Mol. Biol.*, 157:105-132, 1982.