Genome Scale Classification of Extended Proteins by a Predictor SOSUIdumbbell

Nobuyuki Uchikoga[1]
uchikoga@nuap.nagoya-u.ac.jp

Ke Rungcong[1,2]
ke@proteome.bio.tuat.ac.jp

Fumitsugu Akazawa[1]
akazawa@proteome.bio.tuat.ac.jp

Masashi Sonoyama[1]
sonoyama@nuap.nagoya-u.ac.jp

Shigeki Mitaku[1,2]
mitaku@nuap.nagoya-u.ac.jp

[1] School of Engineering, Department of Applied Physics, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Aichi 464-8603, Japan
[2] Department of Biotechnology, Tokyo University of Agriculture and Technology, Naka-cho 2-24-16, Koganei-shi, Tokyo 184-8588, Japan

Keywords: proteome, gene classification, extended protein, DNA binding protein

1 Introduction

Accurate classification of proteins from amino acid sequences is very helpful for the analysis of genome information. There are many novel genes in genomes, which are mainly annotated on the basis of sequence homology. However, many genes in genomes are not homologous to any other genes in the sequence level. Among various bioinformatic methods, the use of physicochemical parameters of amino acid sequences for the classification of proteins has advantages in the analysis of functional unknown sequences [1] because its algorism is not dependent on the sequence homology with any other proteins and understandable in terms of the physical processes of folding.

We have already developed protein classification tools, SOSUI and SOSUIdumbbell [4], which are the predictors of membrane proteins and soluble extended proteins, respectively [3]. We have many data of membrane proteins for testing the accuracy of the methods, and several performance tests have already been reported [2]. However, only a few data are available for extended proteins where 3D-structures are deposited in PDB. For this reason, it was difficult to estimate the accuracy of the classification of extended proteins by SOSUIdumbbell, and the results of the predictions were questionable.

In this work, we have tested the reliability of SOSUIdumbbell by analyzing the unknown sequences, which were predicted as extended proteins. When all amino acid sequences in proteomes were analyzed tens of unknown sequences were classified into the category of extended proteins. Looking at the 3D-structures of the proteins in PDB, which are homologous to those unknown sequences, we found that the analysis by SOSUIdumbbell corrected selects extended proteins.

2 Materials and Methods

The amino acid sequences of complete genomes of all 75 organisms (6 eukaryotes, 57 eubacterias and 12 archaeas) deposited in GenBank/EMBL/DDBJ on January 2002 (231,755 amino acid sequences, including unclearly annotated sequences) were classified in terms of extended-type structure by using SOSUIdumbbell. For results of predicting extended proteins, each amino acid sequence could be assigned to correct positive or false positive (Figure 1). There was no accuracy of the extended structures of predicted genes after searching for extended proteins in genome sequences. Then, amino acid

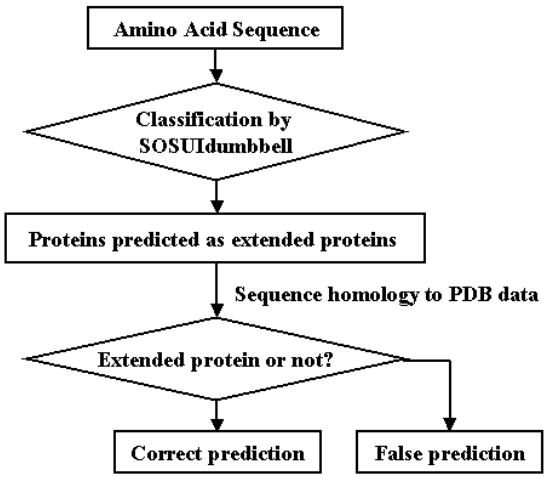

Figure 1: Flow chart of genome scale classification of extended proteins by SOSUIdumbbell.

sequences in PDB database were searched, which were homologous to extended proteins. Particularly, we were interested in functional unknown genes, which had no homologous to any other genes.

3 Results and Discussion

Most of predicted proteins were involved in some regulation processes, which could form complex with other molecules. Most positively charged extended proteins were DNA- or RNA-binding proteins and most of the other charged ones were calmodulin or troponin C. Positive charged extended proteins were more than negative charged ones. Each genome had unknown functional genes occupying almost half of genes classified as extended proteins. Some unknown genes were expected to connect with known functional genes as extended-type proteins.

The process for unknown genes predicted as extended proteins could estimate the accuracy of genome scale classification of extended proteins by SOSUIdumbbell. Functions of most of predicted genes as extended proteins were involved in some regulation, which were, for example, signal transduction, transcription, and so on. Classification of genes in terms of extended-type proteins could help to expand the number of genes related with regulation.

References

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

[2] Mitaku, S., Hirokawa, T., and Tsuji, T., Amphiphilicity index of polar amino acids as an aid in the characterization of amino acid preference at membrane-water interfaces, *Bioinformatics*, 18:608–616, 2002.

[3] Uchikoga, N., Takahashi, S., Ke, R., Akazawa, F., Sonoyama, M., and Mitaku, S., Prediction system for dumbbell-type proteins: SOSUIdumbbell, *Genome Informatics*, 12:328–329, 2001.

[4] `http://sosui.proteome.bio.tuat.ac.jp/sosuidumbbell/dumbbell_submit.html`