A New Method for Characterizing Functionally-Unknown Proteins Using Specific Amino Acid Frequency and Periodicity at the Proteome Level

Kosuke Fujishima[1,2]
t01828kf@sfc.keio.ac.jp

Jun Imoto[1,2]
t01109ji@sfc.keio.ac.jp

Akio Kanai[2]
akio@sfc.keio.ac.jp

Masaru Tomita[1,2]
mt@sfc.keio.ac.jp

[1] Institute for Advanced Biosciences, Keio University, Tsuruoka, Yamagata 997-0017, Japan

[2] Department of Environmental Information, Keio University, 5322 Endo, Fujisawa, Kanagawa 252-8502, Japan

Keywords: proteome, amino acid frequency, amino acid periodicity, archaea, RNA-binding protein

1 Introduction

Complete DNA sequences of over 100 species are determined, and more are to be read. In spite of these results, functional prediction at the proteome level is still in difficulty, because homology-dependant analysis has been the major stream in proteome analysis. In the present study, we developed a new method which identifies a possible protein function based on computational calculation of specific amino acid frequency and periodicity. Using this method, proteins of a hyperthermophilic archaeon, *Pyrococcus furiosus* were classified into several groups by plotting their hydrophobicity score on a graph. As a result, DNA/RNA-binding proteins, ribosomal proteins and membrane proteins were categorized into individual groups. Then by using clustering software, new candidates were chosen for DNA/RNA-binding and ribosomal proteins due to the co-visualization of functionally-known proteins. Moreover, we have cloned the genes for the selected RNA-binding proteins and made the recombinant proteins in *E. coli.* Characterization of RNA-binding activities of the candidates and the efficiency of our method for identifying possible protein functions will be discussed.

2 Dataset and Methods

We used the *P. furiosus* as a model organism because the genome size is relatively small [3] and the gene products exhibit heat-stability so that it is very easy to handle biochemically (2). Amino acid sequences of 2065 *P. furiosus* proteins are prepared from the GenBank database (NCBI genome data archive) and extracted amino acid sequence of 1064 functionally-known proteins including putative gene products were used as the dataset. Then we summed up amino acid hydropathy and molecular weight for each protein were graphed on a distribution plot to figure out whether any functionally similar protein groups would be biased on a specific region. We counted each usage of all 20 amino acids and classified 1064 functionally-known proteins by usage similarity.

We also developed an algorithm to calculate certain periodicity of specific amino acid such as Lysine, Arginine, and Histidine which are negatively charged amino acid indicated as a key component for RNA-binding activity. For each protein, 3 to 28 cycle region of charged amino acid were counted indicating the variety of cycle that each protein possesses. Based on these data, functionally-related proteins were classified using software called Cluster (Dr. Eisen lab. Stanford Univ.). In order to select novel RNA-binding proteins, we have used the same method upon a dataset combining both functionally-known and unknown proteins. For gene cloning and the recombinant protein experiments, we have used the previously described method [1, 4].

3 Results and Discussion

We have developed a new computational analysis method and achieved to found a slight tendency applicable to classify functional proteins only by using primary sequences of amino acids. Fig. 1 strongly suggests that some of the proteins which belong to specific protein families possess a similar usage rate of amino acids and have same chemical characteristic. It has been reported that the major influence of amino-acid composition variability are the hydrophobicity of protein which causing the discrimination of integral membrane protein in *Escherichia coli* [2]. The same feature can be discussed in *P. furiosus*.

From clustering analysis to functionally-known proteins, over 40 DNA/RNA-related proteins were biased in a local region (Fig. 2). By applying the same methods upon functionally-unknown proteins, we have extracted novel candidates for 29 RNA-binding proteins. Three of these proteins (termed PF0029, PF0547 and PF1912) are produced in *E. coli* using molecular biology techniques, and their RNA-binding activities have been examined. The results showed that PF0029 and PF1912 are able to bind a certain RNA stem-loop structure, supporting the usefulness of our method, although their functional specificity is yet to be validated.

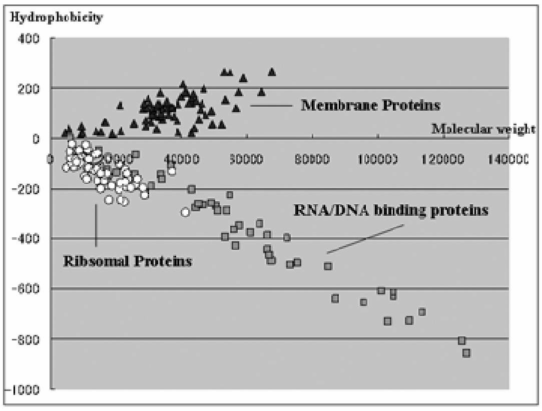

Figure 1: Distribution plot based on hydrophobicity score. Ribosomal proteins, RNA-binding proteins except ribosome proteins, and integral membrane proteins are clearly classified into different regions.

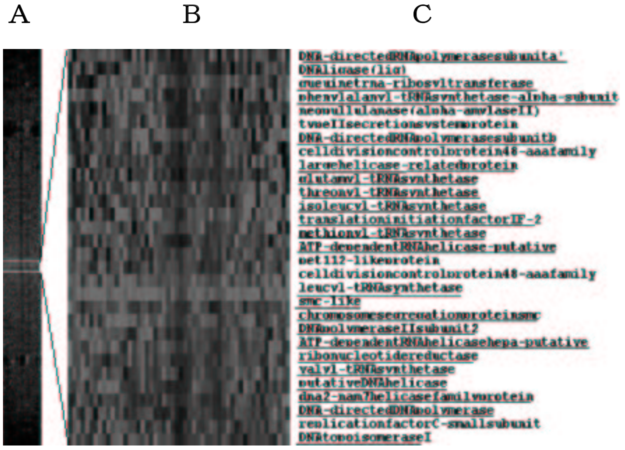

Figure 2: Clustering of 1064 functionally-known proteins in *P. furiosus* by specific amino acid periodicity. **A**: 1064 proteins are classified in line. (A row is the cycle of specific amino acid) **B**: Magnified view of the framed region. **C**: Possible funtion of the responding to the column in B. Underlined proteins are DNA/RNA related.

References

[1] Kanai, A., Oida, H., Matsuura, N., and Doi, H., Expression cloning and characterization of a novel gene that encodes the RNA-binding protein FAU-1 from *Pyrococcus furiosus*, *Biochem J.*, 372(1):253–261, 2003.

[2] Lobry, J.R. and Gautier, C., Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes, *Nucleic Acids Res.*, 22(15):3174–3180, 1994.

[3] Robb, F.T., Maeder, D.L., Brown, J.R., DiRuggiero, J., Stump, M.D., Yeh, R.K., Weiss, R.B., and Dunn, D.M., Genomic sequence of hyperthermophilie, *Pyrococcus furiosus*: implications for physiology and enzymology, *Methods Enzymol.*, 330:134–157, 2001.

[4] Sato, A., Kanai, A., Itaya, M., and Tomita, M., Cooperative regulation for Okazaki fragment processing by RNase HII and FEN-1 purified from a hyperthermophilic archaeon, *Pyrococcus furiosus*, *Biochem. Biophys. Res. Commun.*, 309(1):247–252, 2003.