Distribution of Species Specific Gene Family in Microorganisms

Takeshi Ara [1]
takeshi@bs.aist-nara.ac.jp

Kenji Suzuki [2]
suzuken@kuicr.kyoto-u.ac.jp

Hideo Matuda [13]
matsuda@ics.osaka-u.ac.jp

Hirotada Mori [14]
hmori@gtc.aist-nara.ac.jp

[1] CREST, JST (Japan Science and Technology Corporation)
[2] Division of Molecular Biology and Information, Institute of Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan
[3] Department of Informatics and Mathematical Science, Graduate School of Engineering Science Osaka University, 1-3 Machikaneyama, Toyonaka, Osaka 560-8531, Japan
[4] Research and Education Center for Genetic Information, Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, Nara 630-0101, Japan

1 Introduction

Recently, more than 20 complete genome sequences of microorganisms had been determined, and about fifty thousands open reading frames (ORFs) were predicted. The analysis of these ORFs showed that many orthologous or paralogous ORFs were identified in these genomes [1,2]. The analysis of orthologous or paralogous genes were mainly concentrated on the elucidation of the universal feature through microorganisms so far. Another aspect, diversity, however, is also important objective for understanding microorganisms. Here, we concentrated on the analysis of species specific ORF cluster using 17 species genomes data.

2 Method

38,301 ORFs from 17 complete genomes were clustered by single linkage clustering method using BLAST program (p-value = 1.0e-5), and analyzed more precisely by MDS method [3]. Distribution of species in each clusters were analyzed and the clusters containing ORFs of a single species were extracted. The biological function of these clusters were annotated based on the function annotated members in each clusters.

3 Results and Discussions

9,369 ORFs out of 38,301 total predicted ORFs did not form clusters. These ORFs had no homologs in 17 species based on amino acid sequence similarity. Remaining 28,932 ORFs formed 2,684 clusters. 874 clusters out of 2,684 were constructed by ORFs of only single species. The 874 clusters, containing 2,458 ORFs were identified as species specific paralogous clusters. All species, analyzed in this paper except *Mycoplasma genitalium* were showed to have their own species specific clusters (Fig. 1). Most clusters, 622 out of 874 were formed by only two ORFs. The biggest cluster contained 34 ORFs of *Saccharomyces cerevisiae*, and the function of 3 out of these were annotated as mitochondrial ADP/ATP translocater. In *Escherichia coli*, The biggest ORF cluster contained 14 ORFs and some of them had been annotated as type I pili gene products. The type I pili genes form a gene cluster on the chromosome and they strongly suggested that the duplication events had been occurred in such gene clusters. Such gene duplication might have some advantages for the survival in host organisms.

To investigate the reason why these species specific paralogous ORFs were increased, we analyzed the correlation between the genome size and the number of ORFs contained such clusters. As shown in Fig. 2, the clear correlation between them were observed. The result showed that *Mycobacterium tuberculosis* had clearly more species specific ORFs, on the other hand, *Aquifex aeolicus*, *Borrelia burgdorferi* and *Haemophilus influenzae* had slightly less ORFs than the expected number of such ORFs from the regression line. Even if the genome size were small, the clusters, containing exceptionally large number of ORFs were observed, such as case of *Helicobacter pylori* and *Treponema pallidum*. Elucidation of the advantageness to increase the number of the paralogous genes is an important analysis.

Acknowledgements

This study has been supported by CREST of JST (Japan Science and Technology Corporation) and by a Grant-in-Aid for Scientific Research on Priority Areas "Genome Science" from the Ministry of Education, Science, Sports and Culture in Japan.

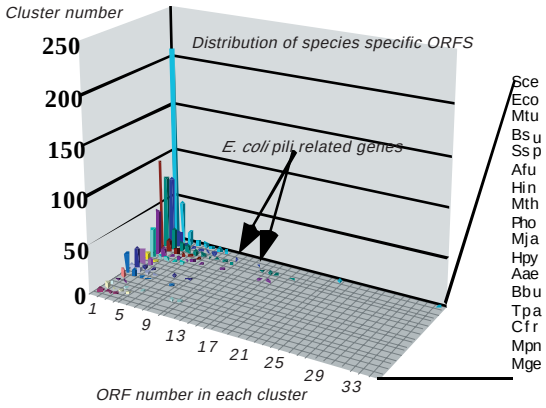

Figure 1: Distribution of species specific ORF.

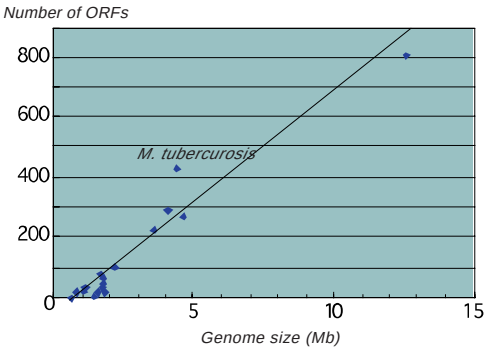

Figure 2: Correlation between the genome size and the clusters cluster number of ORFs contained species specific ORF cluster.

References

[1] Tatusov, R.L., Koonin, E.V., and Lipman, D.J., A genomic perspective on protein families, *Science*, 278:631–637, 1997.

[2] Makarova, K.S., Aravind, L., Galperin, M.Y., Grishin, N.V., Tatusov, R.L., Wolf, Y.I., and Koonin, E.V., Comparative genomics of the archaea (euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell, *Genome Research*, 9:608–628, 1999.

[3] Taketani, H., Tani, H., Matsuda, H., and Hashimoto, A., A method for extracting common conserved regions from amino acid sequences using maximum-density subgraph search algorithm, IPSJ SIG-MPS Report, No.20, 49–54, 1998.