

Automatic Detection of Gene Clusters by P-Quasi Complete Linkage Grouping

Wataru Fujibuchi¹

wataru@kuicr.kyoto-u.ac.jp

Hideo Matsuda²

matsuda@ics.es.osaka-u.ac.jp

Hiroyuki Ogata¹

ogata@kuicr.kyoto-u.ac.jp

Minoru Kanehisa¹

kanehisa@kuicr.kyoto-u.ac.jp

¹ Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan

² Department of Informatics and Mathematical Science, Graduate School of Engineering Science, Osaka University, 1-3 Machikaneyama, Toyonaka, Osaka 560-8531, Japan

1 Introduction

It is known that bacteria genomes are highly organized to express functions cooperatively, for instance, by operons or by locating related genes adjacently on the genome for the transcriptional advantage. Since the complete sequences of several genomes were released, the gene clusters that are defined as locally conserved gene sets beyond species have been progressively studied for these few years [1, 2, 3, 5]. However, it seems difficult to extract functionally important gene clusters manually because the detecting of gene clusters depends on which organisms to compare. Our goal is to extract significant gene clusters automatically with appropriate level of conservation among various species. In our system, tentatively called P-quasi completeness gene cluster detector, or P-CGC detector, the magnitude of significance of gene clusters can be controlled by P-% completeness of linkage of gene clusters, in other words, by the extent of conservation among various species.

2 Method and System

We use 14 microorganisms and the whole processes to detect gene clusters are as follows:

1. Best hit search with Smith-Waterman score equal to or more than 100
2. Extraction of possible gene clusters by SIMIC search, allowing insertion, deletion and permutation of genes
3. P-quasi complete linkage grouping of gene clusters among 14 species
4. Identification of orthologs and paralogs by COG+ method where paralogous genes are added to COGs [4]

3 Results and Discussions

We examined a various set of parameter values, P, and obtained the range of 20–60% that best extracted significant gene clusters. Fig. 1 is an example of extracted gene clusters (P=20%). The ATPase cluster seems to entirely reproduce the manual version of KEGG database. In addition, several interesting characteristics are found in the cluster. Generally speaking, archaeal ATPases contain common subunits, chain β (cog1), α (cog3), B (cog5) and C (cog6) as eubacteria, but they have unique subunits that are different from eubacteria (cog9–12 that apparently substitute for cog0, cog2,

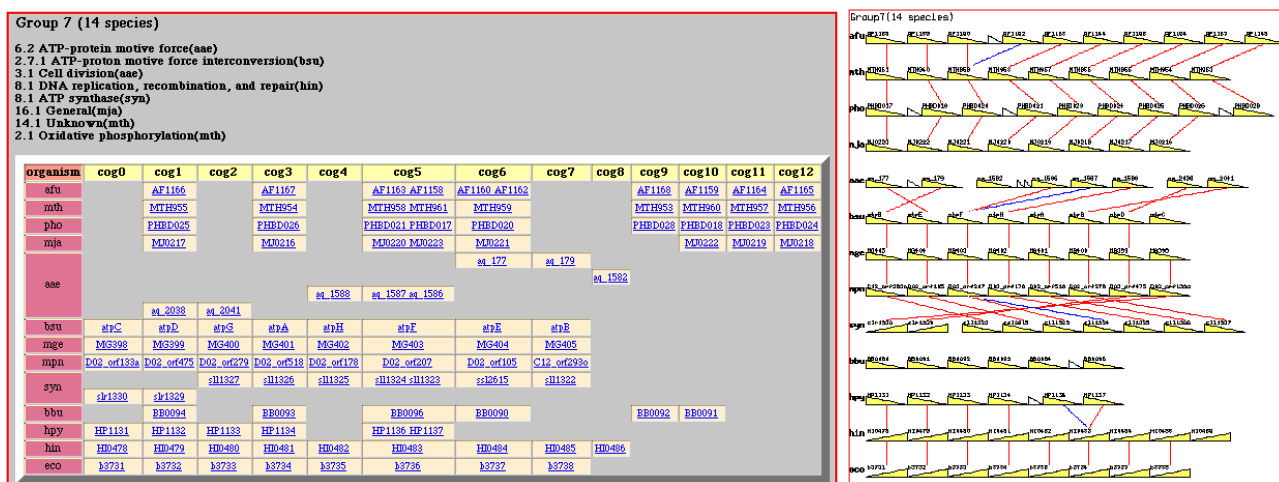


Figure 1: ATPase cluster: table and wired image of orthologs.

cog4, cog7). Moreover, there are duplicated homologs for B chain (cog5). However, it is interesting to find that only *B. burgdorferi* (bbu) has unique archaeal characteristics among eubacteria. Furthermore, in *A. aeolicus* (aae), the cluster is divided into four or more pieces on the genome.

We think that our system can correctly detect appropriate gene clusters automatically and we are now improving and extending it for representing paralogous genes and fusion genes efficiently. Through this work, we expect to analyze the statistical differences of genome organization among organisms and it would provide us with the understanding of the history of genome evolution.

Acknowledgements

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 of Japan. The computation time was provided by the Supercomputer Laboratory, Institute for Chemical Research, Kyoto University.

References

- [1] Mushegian, A.R. and Koonin, E.V., Gene order is not conserved in bacterial evolution, *Trends Genet.*, 12:289–230, 1997.
- [2] Siefert, J.L., Martin, K.A., Abdi, F., Widger, W.R. and Fox, G.E., Conserved gene clusters in bacterial genomes provide further support for the primacy of RNA, *J. Mol. Evol.*, 45:467–472, 1997.
- [3] Tatusov, R.L., Mushegian, A.R., Bork, P., Brown, N.P., Hayes, W.S., Borodovsky, M, Rudd, K.E. and Koonin, E.V., Metabolism and evolution of *Haemophilus influenzae* deduced from a whole-genome comparison with *Escherichia coli*, *Curr. Biol.*, 6:279–291, 1996.
- [4] Tatusov, R.L., Koonin, E.V. and Lipman, D.J., A genomic perspective on protein families, *Science*, 278:631–637, 1997.
- [5] Watanabe, H., Mori, H., Itoh, T. and Gojobori, T., Genome plasticity as a paradigm of eubacteria evolution, *J. Mol. Evol.*, 44(Supple 1):S57–S64, 1997.