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

Wataru Fujibuchi [1]
wataru@kuicr.kyoto-u.ac.jp
Hiroyuki Ogata [1]
ogata@kuicr.kyoto-u.ac.jp
Hideo Matsuda [2]
matsuda@ics.es.osaka-u.ac.jp
Minoru Kanehisa [1]
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.

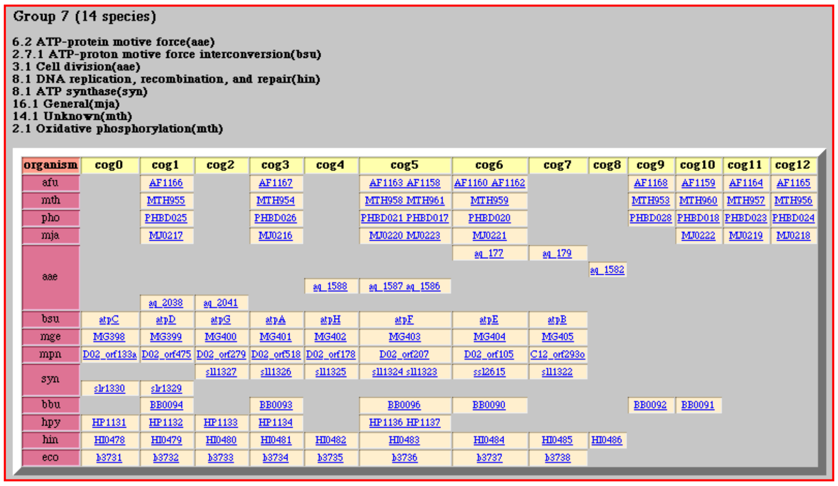
2 Method and System

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,



Group 7 (14 species)

6.2 ATP-protein motive force(aae)
2.7.1 ATP-proton motive force interconversion(bsu)
3.1 Cell division(aae)
8.1 DNA replication, recombination, and repair(hin)
8.1 ATP synthase(syn)
16.1 General(mja)
14.1 Unknown(mth)
2.1 Oxidative phosphorylation(mth)

organism	cog0	cog1	cog2	cog3	cog4	cog5	cog6	cog7	cog8	cog9	cog10	cog11	cog12
afu		AF1166		AF1167		AF1163 AF1158	AF1160 AF1162			AF1168	AF1159	AF1164	AF1165
mth		MTH955		MTH954		MTH958 MTH961	MTH959			MTH953	MTH960	MTH957	MTH956
pho		PHBD025		PHBD026		PHBD021 PHBD017	PHBD020			PHBD028	PHBD018	PHBD023	PHBD024
mja		MJ0217		MJ0216		MJ0220 MJ0223	MJ0221				MJ0222	MJ0219	MJ0218
aae							aq_177	aq_179					
									aq_1582				
					aq_1588	aq_1587 aq_1586							
		aq_2038	aq_2041										
bsu	atpC	atpD	atpG	atpA	atpH	atpF	atpE	atpB					
mge	MG398	MG399	MG400	MG401	MG402	MG403	MG404	MG405					
mpn	D02_orf133a	D02_orf475	D02_orf279	D02_orf518	D02_orf178	D02_orf207	D02_orf105	C12_orf293o					
syn			sll1327	sll1326	sll1325	sll1324 sll1323	ssl2615	sll1322					
	slr1330	slr1329											
bbu		BB0094		BB0093		BB0096	BB0090			BB0092	BB0091		
hpy	HP1131	HP1132	HP1133	HP1134		HP1136 HP1137							
hin	HI0478	HI0479	HI0480	HI0481	HI0482	HI0483	HI0484	HI0485	HI0486				
eco	b3731	b3732	b3733	b3734	b3735	b3736	b3737	b3738					

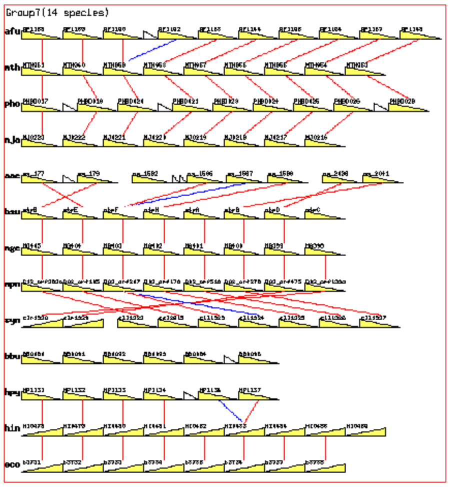


Figure 1: ATPase cluster: table and wired image of orthologs.

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.