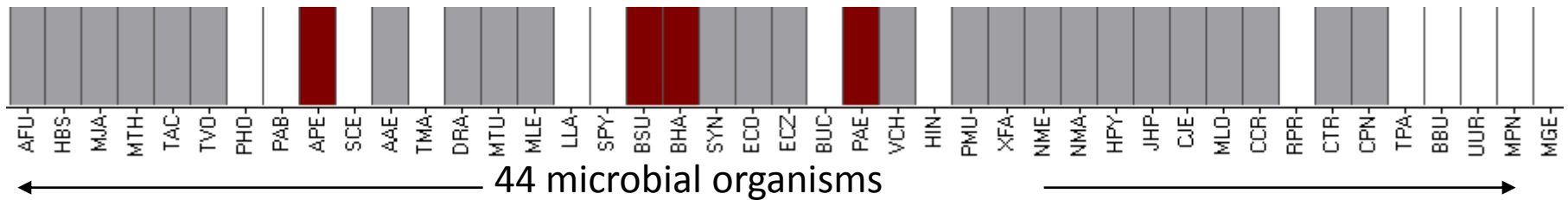


Tutorial No. 7

Gene duplications

Clusters of orthologous groups of genes = gene families

Glutamate-1-semialdehyde aminotransferase (COG0001)



Gene not present in the organism



Gene present in the organism



Gene present in more than one copy in the organism

COG database

<http://www.ncbi.nlm.nih.gov/COG/>

Clusters of **O**rthologous **G**roups of proteins (COGs) were delineated by comparing protein sequences encoded in complete genomes, representing major phylogenetic lineages.

Each COG consists of individual proteins or groups of paralogs from at least 3 lineages and thus corresponds to an ancient conserved domain.

cog_prof_mat_whog matrix

	Tvo 50339	Tac 2303	Mka 190192	Mja 2190	Mth 145262	Mac 188937	Hbs 64091	Afu 2234	
Pho 53953	Archaea	Archaea	Archaea	Archaea	Archaea	Archaea	Archaea	Archaea	
Archaea	Thermoplasma	Thermoplasma	Methanopyrus	Methanococcus	Methanohalobacter	Methanosarcina	Halobacterium	Archaeoglobus	
Pyrococcus	ma horikoshii	ma acidophilum	us kandleri	ccus jannaschii	thermautotrophicus	acetivorans str.C2A	um sp. NRC-1	bus fulgidus	
0	1	1	1	1	1	1	1	1	COG0001
1	0	0	1	1	1	1	0	1	COG0002
0	2	1	1	1	1	0	2	0	COG0003
0	1	2	1	2	2	3	0	3	COG0004
2	1	1	1	1	1	1	2	1	COG0005
3	1	2	1	1	1	2	2	1	COG0006
0	1	1	1	1	1	1	1	2	COG0007
1	1	1	1	1	1	1	1	1	COG0008
1	1	1	1	1	1	1	1	1	COG0009
2	1	1	1	1	1	1	2	1	COG0010
1	1	1	1	1	1	1	1	1	COG0011
1	1	1	1	1	1	1	2	1	COG0012
1	1	1	1	1	1	1	1	1	COG0013
0	0	0	0	0	0	0	0	0	COG0014
1	1	1	1	1	1	1	1	1	COG0015
1	1	1	1	1	1	1	1	1	COG0016
3	3	3	1	1	1	1	1	1	COG0017
1	1	1	1	1	1	1	1	1	COG0018
0	1	1	1	1	1	0	1	1	COG0019
1	1	1	1	1	1	2	2	2	COG0020

Name of microorganism

Afu 2234
Archaea
Archaeoglobus fulgidus

COG0001

COG0002

COG0003

3

COG0004

COG0005

COG0006

COG0007

COG0008

COG0009

COG0010

COG0011

COG0012

COG0013

COG0014

COG0015

COG0016

COG0017

COG0018

COG0019

COG0020

Name of gene family

No. of representative genes in family
COG004 in organism Mac 188937

Problem Set 7

- **Problem 1:** Show the distribution of gene family sizes for your genome.
- **Problem 2:** Identify the biggest (or one of the biggest) paralog family (COG) in your genome. What is its size distribution across the other genomes?
- **Problem 3:** How many of your genome's genes are paralogs (have duplicates in the genome assigned to the same gene family)?
- **Problem 4:** What fraction of your genome's genes belong to families with representatives across all other genomes?
- **Problem 5:** In comparison with the genome next to yours in the matrix (your column + 1), which families are significantly larger (more than 3 more members) and which are significantly smaller (more than three less, yet still present)?