**A NATURAL TAXONOMY OF GENE FAMILIES FROM COMPLETE GENOMES.**

TATUSOV R.L.+, KOONIN E.V., LIPMAN D.J.

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda MD 20894, USA;
e-mail: tatusov@ncbi.nlm.nih.gov, e-mail: koonin@ncbi.nlm.nih.gov, e-mail: lipman@ncbi.nlm.nih.gov,

+Corresponding author

Keywords: genome sequences, orthologs, paralogs, sequence comparison, cluster of orthologous groups

To extract maximum information from the rapidly accumulating genome sequences, all conserved genes need to be classified according to their homologous relationships. By comparing proteins encoded in 7 complete genomes from 5 major phylogenetic lineages and elucidating consistent patterns of sequence similarities, we delineated 720 Clusters of Orthologous Groups (COGs). Each COG consists of individual, orthologous proteins or orthologous sets of paralogs from at least three lineages. Orthologs typically have the same function, allowing transfer of functional information from one member to an entire COG. This automatically yields a number of functional predictions for poorly characterized genomes. The COGs comprise a framework for functional and evolutionary genome analysis.

Knowing the inventory of conserved genes responsible for housekeeping functions and understanding the differences in the genetic basis of these functions in different phylogenetic lineages is central to understanding life itself, at least at the level of a single cell. Complete sequences are indispensable for achieving this goal, as it is the only type of information that can be used to delineate the complete network of relationships between genes from different genomes. Furthermore, only with complete genome sequences, it is possible to ascertain that a particular protein implicated in an essential function is not encoded in a given genome. Accordingly, an alternative protein for the respective function should be sought among the functionally unassigned gene products (1-11). With multiple genome sequences, it is possible to delineate protein families that are highly conserved in one domain of life but are missing in the others. Such information may be critically important – for example, the families that are conserved among bacteria but are missing in eukaryotaes comprise the pool of potential new targets for broad-spectrum antibiotics.

The knowledge of all the gene sequences from multiple complete genomes redefines the problem of gene classification. It becomes feasible to replace more or less arbitrary clustering of genes by similarity with a complete, consistent system, in which the groups are likely to have evolved from a single ancestral gene. Such a natural classification of genes will provide a framework for evolutionary studies and for rapid, largely automatic functional annotation of newly sequenced genomes.

The relationships between genes from different genomes are naturally represented as a system of homologous families that include both orthologs and paralogs. Orthologs are genes in different species that evolved from a common ancestral gene by speciation; by contrast, paralogs are genes related by duplication within a genome (12). Normally, orthologs retain the same function in the course of evolution, whereas paralogs evolve new functions, even if related to the original one. Thus identification of orthologs is critical for reliable prediction of gene functions in newly sequenced genomes. A naive operational definition would simply maintain that for a given gene from one genome, the gene from another genome with the highest sequence similarity is the ortholog. Given the complete genome sequences, this straightforward approach often gives credible results, especially when the compared species are not too distant phylogenetically (13). At larger phylogenetic distances, however, the situation becomes more complicated. Firstly, if gene duplications occurred in each of the given two clades subsequent to their divergence, only a many-to-many relationship will adequately describe orthologs, and accordingly, detection of the highest similarity will not result in the identification of the complete set of orthologs. Secondly, when the best hit is not highly significant statistically, which is common in the case of phylogenetically distant relationships (14), it simply may be spurious. On the other hand, attempts to apply a restrictive similarity cut-off are likely to result in a number of orthologs being missed.

Given the existence of one-to-many and many-to-many orthologous relationships, we redefined the task of identification of orthologs as delineation of Clusters of Orthologous Groups (COGs). Each COG consists of individual orthologous genes or orthologous groups of paralogs from three or more phylogenetic lineages. Each COG is assumed to have evolved from an individual ancestral gene through a series of speciation and duplication events.

In order to delineate the COGs, all pairwise sequence comparisons among the proteins encoded in the complete genomes were performed using the BLASTPGP program (15). For each protein, the best hit (BeT) in each of the other genomes was detected. The identification of COGs was based on consistent patterns in the graph of BeTs. The simplest and most important of such patterns is a triangle, which typically consists of orthologs. The consistency between BeTs resulting in triangles does not depend on the absolute level of similarity between the compared proteins and thus allows the detection of orthologs among both slowly and fast evolving genes. The algorithm further included verification that the BeTs included in a triangle formed a consistent multiple alignment; triangles that did not contain a conserved motif were disregarded.

The described analysis resulted in 720 COGs, including 6814 proteins and distinct domains. Most of the COGs are relatively small groups of proteins. One third of the COGs (240 COGs with 1406 proteins) contain one representative of each of the included species (no paralogs), and 192 more COGs include paralogs from only one species, most frequently yeast (87 COGs). A notable aspect of many COGs is the differential behavior of paralogs. It is typical that one of the paralogs, e.g. in yeast, shows consistently higher similarity to the orthologs in all or most of the other species.

For the majority of the COGs, the protein function is either known from direct experiments, mainly in E. coli or yeast, or can be confidently inferred on the basis of significant sequence similarity to functionally characterized proteins from other species. It has to be emphasized that construction of the COGs includes automatic prediction of the function for numerous genes, particularly from the poorly characterized genomes such as M. jannaschii. There is, however, a substantial fraction of the COGs (about 10%), for which only general functional prediction, typically of biochemical activity, but not the actual cellular role could be made, and for another 5%, there was no functional clue (Table 1). Each of the COGs includes proteins from at least 3 major clades whose divergence time is estimated to be over a billion years (16), i.e. they all are ancient, conserved families with important, cellular functions. Therefore the proteins belonging to the "mysterious" COGs are good candidates for directed experimental studies.

Table 1. Phylogenetic patterns

**Bacteria +**

 **Eukarya + Eukarya + Archaea + Bacteria only**

 **Archaea Bacteria Bacteria**

 **pattern # pattern # pattern # pattern #**

 **eh--cmy 122 eh--c-y 79 eh--cm- 52 ehgpc-- 54**

 **ehgpcmy 115 ehgpc-y 65 e---cm- 43 e-gpc-- 4**

 **e---cmy 37 e---c-y 56 ehgpcm- 15 eh-pc-- 2**

 **eh---my 17 ehgp--y 5 e-gpcm- 4**

 **----cmy 13 e-gpc-y 2 -h--cm- 3**

 **e----my 7 -h--c-y 1 eh-p-m- 2**

 **--gpcmy 4 eh-pc-y 1 ehgp-m- 2**

 **eh-p-my 2 --gpc-y 1 --gpcm- 1**

 **ehgp-my 2 -hgp--y 1 e-gp-m- 1**

 **-h---my 2 e-gp--y 1**

 **e-gpcmy 2 e--p--y 1**

 **--gp-my 1**

 **eh-pcmy 1**

 **e-gp-my 1**

 **14 326 11 213 9 123 3 60**

 **(45%) (29%) (17%) (8%)**

The phylogenetic distribution of COG membership can be conveniently presented in terms of "phylogenetic patterns" that show the presence or absence of each of the analyzed species (Table 1). The two most abundant patterns easily could be predicted – all species ("ehgpcmy") and all species except for the mycoplasmas ("eh\_\_cmy"). What appears much less trivial, is that these patterns together encompass only one third of all COGs. This emphasizes the remarkable fluidity of genomes in evolution, revealed in spite of the fact that the analysis concentrated on ancient conserved families. Multiple solutions for the same important cellular function appear to be a rule rather than an exception, at least when phylogenetically distant species are considered (13-14). On the other hand, it is notable that the 8 most frequent patterns, which together account for 85% of the COGs (Table 1), all include both E. coli and Synechocystis, emphasizing the congruency between these genomes.

The distribution of the COGs by the three domains of life, with only 45% of the COG including representatives of bacteria, archaea, and eukarya, is another manifestation of the dynamics of gene families in evolution (Table 1). The unusual, rare patterns are of particular interest, suggesting the possibility of unexpected findings.

The COG system allows automatic functional and phylogenetic annotation of genes and gene sets. Sequences can be submitted for searching at http://www.ncbi.nlm.nih.gov/COG/cognitor.html. Similarly to the procedure used for the construction of the COGs, the criterion for adding likely orthologs from other genomes to the COGs is based on the consistency between the observed relationships. A protein is compared to the database of protein sequences from complete genomes and is included in a COG if at least two BeTs fall into it. Given that the COGs were constructed from proteins encoded in complete genomes, it is not a requirement that newly included proteins also originate from a complete genome. Indeed, while the unsequenced portion of a genome may encode proteins with the highest similarity to those included in COGs, the BeTs will not change for the products of already sequenced genes.

The COGs bring together the fields of comparative genomics and protein classification. Among numerous possible approaches to protein classification, the COGs appear to be unique as a prototype of a natural system, which has as its basic unit a group of descendants of a single ancestral gene. Typically, such a group is associated with a conserved, specific function, so that when a new protein is included in a COG, this automatically entails functional prediction.

With the forthcoming flood of genome sequences, a coherent framework for understanding these genomes from both the functional and evolutionary viewpoints is a must. We regard the current collection of COGs as a crude first version of such a framework. Inclusion of new, phylogenetically diverse genomes as well as further development of the procedures used to derive and analyze COGs will hopefully result in refinement of this system, making it a solid platform for genome annotation and evolutionary genomics.

**Acknowledgements**

We thank Roland Walker, Hidemi Watanabe, Michael Galperin, Kira Makarova, and Michael Rozanov for valuable help with data analysis; Kenn Rudd, Tyra Wolfsberg, and David Landsman for unpublished data; and Peer Bork, Michael Gelfand, Michael Roytberg, Michael Rozanov, and Roland Walker for helpful discussions.

**References**

1. R. D. Fleischmann et al., Science 269, 496 (1995).
2. C. M. Fraser et al., ibid 270, 397 (1995);
3. R. Himmelreich et al., Nucleic Acids Res. 24, 4420 (1996).
4. T. Kaneko et al., DNA Res. 3, 109 (1996).
5. F. R. Blattner et al., Science 277, 1453 (1997).
6. C. J. Bult et al. ibid. 273, 1058 (1996).
7. Goffeau et al. ibid 274, 546 (1996).
8. H. W. Mewes et al. Nature 387, 7 (1997).
9. C. R. Woese, Curr. Biol. 6, 1060 (1996).
10. E. V. Koonin, Genome Res.7, 418 (1997).
11. E. V. Koonin, A. R. Mushegian, K. E. Rudd, Curr. Biol. 6, 404 (1996).
12. W. M. Fitch, Syst. Zool. 19, 99 (1970).
13. R. L. Tatusov et al., Curr Biol 6, 279 (1996).
14. E. V. Koonin, A. R. Mushegian, M. Y. Galperin, D. R. Walker, Molec. Microbiol. 25, 619.
15. S. F. Altschul et al. Nucleic Acids Res. 25, 3389 (1997).
16. R. F. Doolittle et al. Science 271, 470 (1996).