

An Evolution-Based Clustering Method to Separate Orthologous Genes from Out-paralogs*

Raja Jothi[†], Elena Zotenko[†] Teresa M. Przytycka^{†‡}

Keywords: Orthologs, Paralogs, Clusters of Orthologous Groups (COGs), Clustering

1 Introduction

Two genes from two different species are said to be *orthologs* if they evolved directly from a single gene in the last common ancestor [1]. Genes that evolved from a single gene that was duplicated within a genome are called *paralogs* [1]. Paralogs are further classified into two types: *in-paralogs* and *out-paralogs* [2, 3]. In-paralogs are paralogs that were duplicated after the speciation event, and out-paralogs are paralogs that were duplicated before the speciation event. Fig. 1 illustrates the difference between in-paralogs and out-paralogs. Typically, orthologs perform the same function, whereas a paralog in a genome evolves to perform a new function.

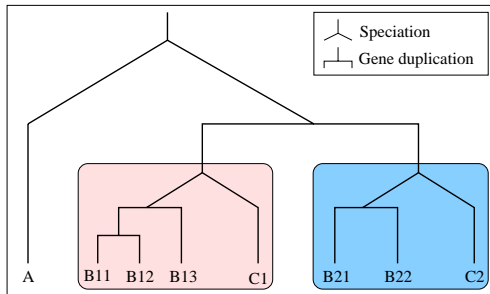


Figure 1: There are two speciation events (inverse Y junctions), and three gene-duplication events (inverse T junctions). Two genes whose common ancestor is at a Y junction are orthologous, e.g., A and B11, and B21 and C2. Two genes whose common ancestor is at a T junction are paralogous, e.g., B11 and B13, and B12 and C2. B11, B12, and B13 are in-paralogs to A (or C) because the speciation event occurred before the B1* duplication event, whereas B1* and C1 are out-paralogs to B2* and C2 as the duplication occurred before speciation.

Tatusov et al. [4] developed Clusters of Orthologous Groups (COGs), a classification of genes across multiple complete genomes based on their homologous relationship. A COG comprises individual orthologous genes or orthologous groups of paralogs (in-paralogs and out-paralogs) from three or more lineages. COGs does not distinguish between in-paralogs and out-paralogs. Since the idea of COGs present a framework using which gene functions for newly sequenced or poorly characterized genomes can be predicted, it is very important that COGs contain only orthologs (and in-paralogs), and not out-paralogs. In this paper, we present a new evolution-based iterative clustering method, which can be used to identify out-paralogs in a given COG and thus refine the COG to contain only orthologs and in-paralogs. Through repeated clustering of a COG, one can construct a gene tree that shows not only the speciation, but also the duplication events. Our clustering method is simple and can easily be automated and used to classify any phylogenetically related set of entities. Previously, Remm et al. [3] presented a method that clusters orthologs and in-paralogs. However, their method can only find orthologous genes between any two species, and not across several species.

* This work was supported by the intramural research program of the National Institutes of Health.

[†] National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, MD 20894, USA.

[‡] Corresponding author. Email: przytyck@mail.nih.gov

2 Materials and Methods

Let N be the number of proteins in a given COG. Sequences contained in a given COG were aligned using CLUSTALW1.83. Similarity matrix from the multiple sequence alignment were calculated using CLUSTALW. Let I denote the vector of evolutionary distances in the similarity matrix from protein i . A new matrix called the correlation coefficient matrix was calculated, in which each entry $-1.0 < s_{ij} < 1.0$ represents the correlation coefficient between column vectors I and J . Clustering is performed by constructing a graph with N vertices (each representing a protein in the COG), in which an edge exists between vertices i and j if and only if s_{ij} is non-negative. The resulting graph may or may not be fully connected. Based on the number of edges crossing a cut between two connected components (subgraphs) in the graph, the graph may qualify to be split into two subgraphs, with the subgraph containing proteins from majority of the species (in the original COG) denoting the new refined COG, while the other subgraph denoting the set of out-paralogs with respect to the refined COG. Our clustering technique breaks down a COG into at most two clusters. One may choose to refine the COG further by performing the above steps yet again, starting from the alignment of sequences. Checks and balances are provided to prevent the breaking-up of a given COG along kingdom lines.

3 Results

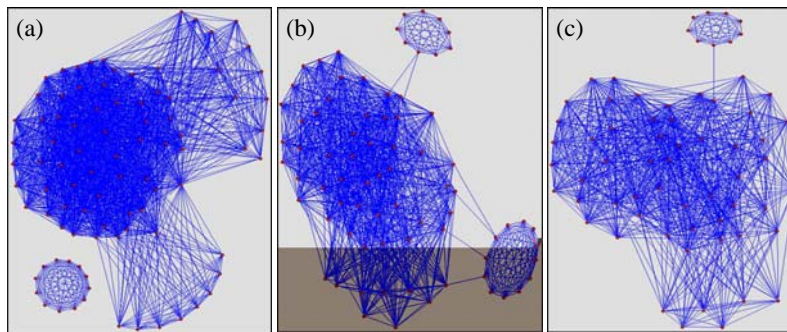


Figure 2: (a) Graph representing COG0616 with 89 proteins from 51 species. Two nodes (proteins) are connected by an edge if and only if their distance vectors to other nodes (including themselves) are positively correlated. Clustering shows that the 12-node connected component does not belong to the COG, indicating that these 12 proteins are out-paralogs to the proteins in the main cluster, leaving the COG with 77 proteins from 47 species. (b) Second round of clustering is applied to the refined COG0616, and 12 more out-paralogs are removed, leaving the COG with 65 proteins from 46 species. (c) Third round of clustering removes 9 more out-paralogs, leaving the COG with 56 proteins from 46 species. Further clustering is not recommended as it only breaks the COG along kingdom lines.

References

- [1] Fitch, W. M. (1970). Distinguishing homologous from analogous proteins. *Syst. Zool.*, 19:99–113.
- [2] Koonin, E. V. and Sonnhammer, E. L. (2002). Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet.*, 18(12):619–620.
- [3] Remm, M. and Storm, C. E. and Sonnhammer, E. L. (2001). Automatic Clustering of Orthologs and In-paralogs from Pairwise Species Comparisons. *J. Mol. Biol.*, 314:1041–1052.
- [4] Tatusov, R. L. and Koonin, E. V. and Lipman, D. J. (1997). A Genomic Perspective on Protein Families. *Science*, 278:631–637.