Hierarchical Clustering Procedure for Grouping Orthologous Domains in Multiple Genomes

Ikuo Uchiyama
uchiyama@nibb.ac.jp
National Institute for Basic Biology, Okazaki National Research Institutes,
Nishigonaka 38, Myodaiji, Okazaki 444-8585, Japan

Keywords: ortholog identification, clustering, homologous domains

1 Introduction

Identification of orthologous genes, which are defined as homologous genes derived from speciation in the history of evolution [1], is a crucial step for comparative analysis of multiple genomes. Basically, evidence to identify orthologs includes: *i*) highest level of similarity, *ii*) conservation of gene arrangement on the chromosome, *iii*) consistency between gene phylogeny and species phylogeny. However, because of very large evolutionary distances as well as complex evolutionary events such as genome rearrangements and horizontal gene transfers, ortholog identification among various microbial genomes is not so trivial task. Many of the previous works (e.g. [2, 5]) mainly relied on the best hit relationships among all-against-all protein sequence comparisons between two genomes, but generally they required several additional procedures such as domain splitting, gene order comparison and phylogenetic consideration. On the other hand, numerous clustering procedures have been developed for homology domain identification (e.g. [4]), but generally they are not suitable for ortholog grouping.

Here, I propose a hierarchical clustering approach which is flexible enough to incorporate many of the above ortholog criteria. Especially, the procedure is suitable for splitting genes into homologous domains minimally required for ortholog grouping.

2 Overview of the Method

The procedure takes as input results of all-against-all protein sequence similarity search including similarity scores and matched regions, and at first constructs a graph of similarity relationships (Fig. 1 A). The clustering procedure is basically successive contraction of this graph based on UPGMA [3]. In each iteration, two nodes, or operational taxonomic units (OTUs) [3], connected with the best similarity relationship are merged and similarity score on each merged edge is assigned averaged score of all relationships contained in it. When some OTUs are connected only one of the best scoring OTUs, a fixed score (parameter) worse than a given clustering cutoff is assigned for the missing relationship. The procedure is repeated until the best similarity score becomes worse than the cutoff.

This basic procedure works well for single domain proteins but not for multi-domain proteins. To treat multi-domain proteins, the procedure is further extended (Fig. 1). In each iteration, merged OTU is split into at most 5 nodes according to the matched region of the best scoring edge: the matched region itself and left and right overhangs of the first and the second OTUs. Edges connected to the merged OTUs are appropriately reconnected to one or more of the new OTUs according to the matched regions of them (Fig. 1 B), and matched lengthes as well as scores of merged edges connected to the merged OTU are averaged over all relationships contained in them. In addition, the nodes newly created are connected each other by neighboring links (broken lines in Fig. 1 B), which are also merged through the clustering process.

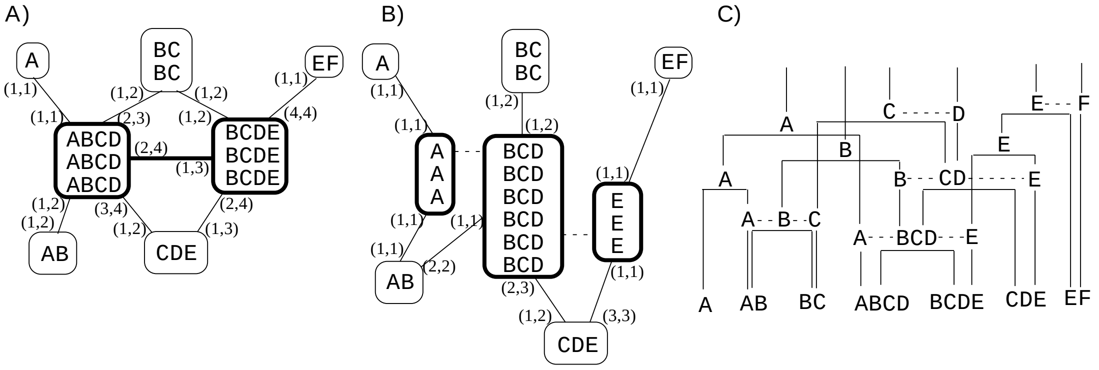

Figure 1: Basic strategy. A,B) Similarity graph before and after merging two nodes (OTUs) connected with the best similarity relationship (bold line). The alphabets A ∼ F denote (priorly unknown) homologous domains and the numbers in parentheses are the beginning and ending positions of the matched regions. Broken lines indicate neighboring links. C) Resultant hierarchical clustering graph.

Resultant structure created by this procedure is a hierarchical clustering graph that consists of overlapping trees as shown in Fig. 1 C. From this structure, one can easily identify homologous clusters in domain level by traversing it from top to bottom. Moreover, the history of domain splitting of each gene can be traced by traversing it in reverse direction. Thus, one can split genes into domains minimally required for ortholog grouping, by specifying some appropriate internal nodes as roots. Root node of each orthologous cluster is determined by another procedure, phylogenetic-cut, where each original root node is recursively cut until two clusters merged at that node share no or little intraspecies paralogous genes. Note that the procedure has also ability to incorporate the information of gene orders when neighboring genes are priorly connected with links and appropriate scoring scheme is introduced, although details are now under investigation.

Of course, obtained structure like Fig. 1 C does not always coincide with the exact evolutionary history, nevertheless in many cases classification is well consistent with the previous ones like COG [5]. Moreover, the procedure is rapid enough to be executed repeatedly with changes of parameters or datasets. Thus, it is useful for the task of grouping a large number of genes in multiple genomes, when combined with an appropriate gene classification workbench such as MBGD [6].

References

[1] Fitch, W.M., Distinguishing homologous from analogous proteins, *Syst. Zool.*, 19(2):99–113, 1970.

[2] Huynen, M.A. and Bork, P., Measuring genome evolution, *Proc. Natl. Acad. Sci. USA*, 95(11):5849–5856, 1998.

[3] Sneath, P.H.A. and Sokal, R.R., *Numerical Taxonomy*, Freeman, San Francisco, 1973.

[4] Sonnhammer, E.L. and Kahn, D., Modular arrangement of proteins as inferred from analysis of homology, *Protein Sci.*, 3:482–492, 1994.

[5] Tatusov, R.L., Galperin, M.Y., Natale, D.A. and Koonin, E.V. The COG database: a tool for genome-scale analysis of protein functions and evolution, *Nucleic Acids Res.*, 28:33–36, 2000.

[6] http://mbgd.genome.ad.jp/.