COCO-CL: Hierarchical Clustering of Homology Relations Based on Evolutionary Correlations (Supplementary Material)

Raja Jothi, Elena Zotenko, Asba Tasneem, Teresa M. Przytycka

COCO-CL on raw BLAST searches

To test the stand-alone operability of COCO-CL, we constructed clusters by including the hits from BLAST queries, whose e-values are less than or equal to 1e-5 and alignment overlap (or the sequence identity) with the query protein is at least above a certain threshold. For the first test case, we generated a set of proteins by BLAST-ing *Homo sapiens* protein gi:558586, which is a glutaminyl-tRNA synthetase. There were 771 hits (from the nr database) with e-value $\leq$1e-5, most of which are glutaminyl- and glutamyl-tRNA synthetases with a few glutamine-, glutamate-, and asparagine-tRNA ligases. Imposing alignment overlap of 75%, 67%, 60%, and 50% reduced the number of hits to 53, 259, 289, and 308 proteins, respectively. Separately, imposing sequence identity of 33%, 40%, and 50% reduced the number of hits from 771 to 51, 256, and 450 proteins, respectively. It is well-accepted that a conservative approach with at least 75% alignment overlap or a significant sequence identity results in a set of proteins that are truly orthologous. Such an approach ensures that the resulting set of proteins do not contain protein fragments or multi-domain proteins sharing only a subset of domains with the query protein. On the other hand, a BLAST search with such stringent constraint(s) may miss out on some orthologous proteins. In our case, we used an e-value cutoff of 1e-5 along with alignment overlap of at least 60% to obtain a set of 289 proteins from 220 organisms. Applying COCO-CL on this set resulted in a split with clustering bootstrap score 0.61, well below the set threshold 0.75. Ignoring the low bootstrap score, COCO-CL would have called this split as a speciation event as the set of organisms on either side of the split are different, thus making the confidence score $\sigma = 0$. On manual examination, almost all of the 289 proteins are truly orthologous to the query protein gi:558586.

While 60% alignment overlap seems to be a good threshold for the case above, it is not that good of a threshold for the ``uracil phosphoribisyltransferase" family. On BLAST-ing *E.coli* protein gi:16130423 from this family, we obtained a set of 367 proteins from 290 organisms with tens of ``unnamed or hypothetical" proteins. Since COCO-CL, in each run, clusters the input set into exactly two subclusters with zero or negative correlation, the amount of ``noise" in the initial set to be clustered has an influence on the clustering process. Since the set we generated using BLAST contains numerous hypothetical proteins, COCO-CL did not yield desired clustering results. However, when the alignment overlap threshold was reduced to 70%, the size of the input set reduced to 327 proteins from 280 organisms. COCO-CL split this reduced set of proteins two subclusters with clustering bootstrap score 0.62. COCO-CL correctly did not split this set of proteins as, on manual verification, they are all orthologous, annotated as ``uracil phosphoribisyltransferase." An approach to cluster BLAST hits of a single query (with just the e-value cutoff of 1e-5) using COCO-CL may not produce correct orthology results without imposing additional constraints such as a significant alignment overlap, sequence identity/similarity, bit-scores, or a combination thereof. Setting these constraints may not be straightforward as one may have to

take into account the scenarios that may arise while BLAST-ing multi-domain proteins. In summary, our tests indicate that applying COCO-CL on raw BLAST searches may not produce correct clustering results or speciation/duplication predictions. Even though the COCO-CL results on these two cases seem to be consistent with what is known, it is not quite clear as to what restrictions one needs to place on raw BLAST searches so that the resulting hits are clean-enough for COCO-CL to be applied upon.

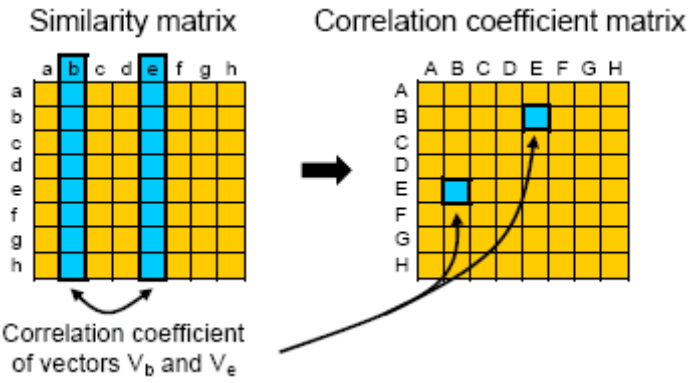


Fig 1: Computing correlation coefficient matrix from a similarity matrix. Each entry r_{ij} in the correlation matrix represents the agreement between column vectors V_i and V_j in the similarity matrix. Pearson's correlation coefficient is used to measure the agreement between column vectors.

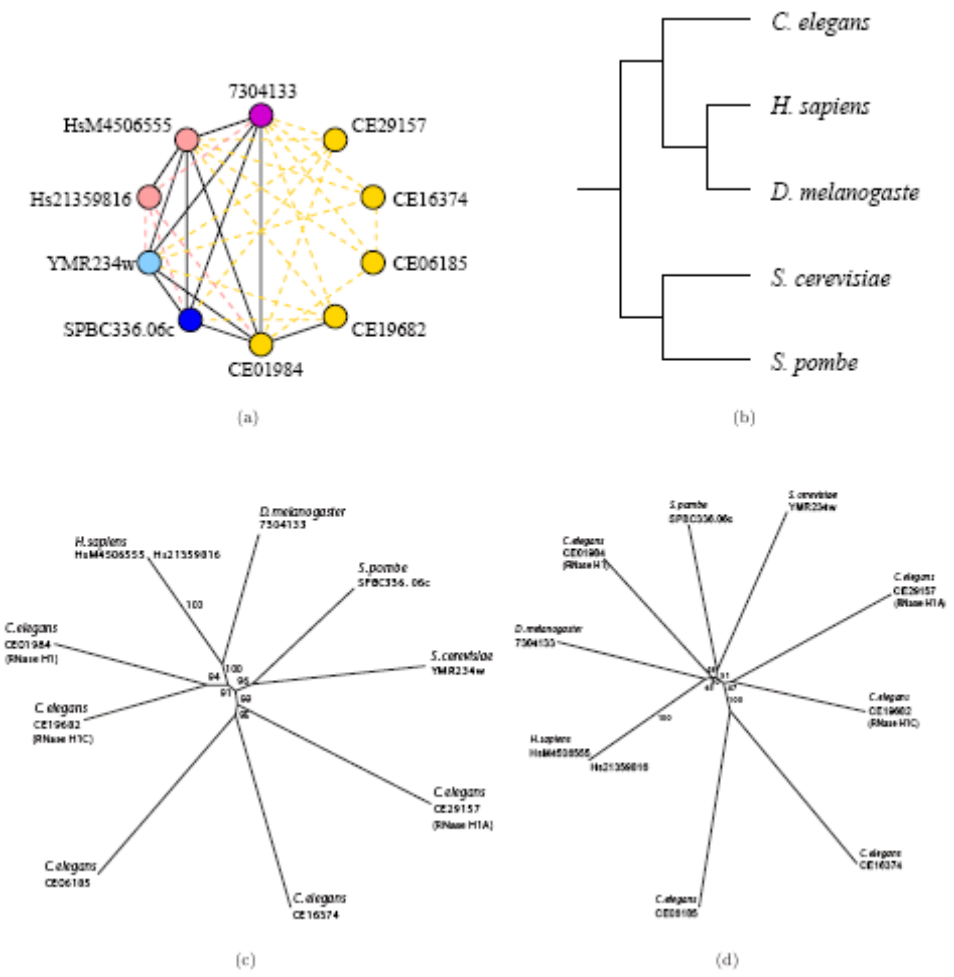
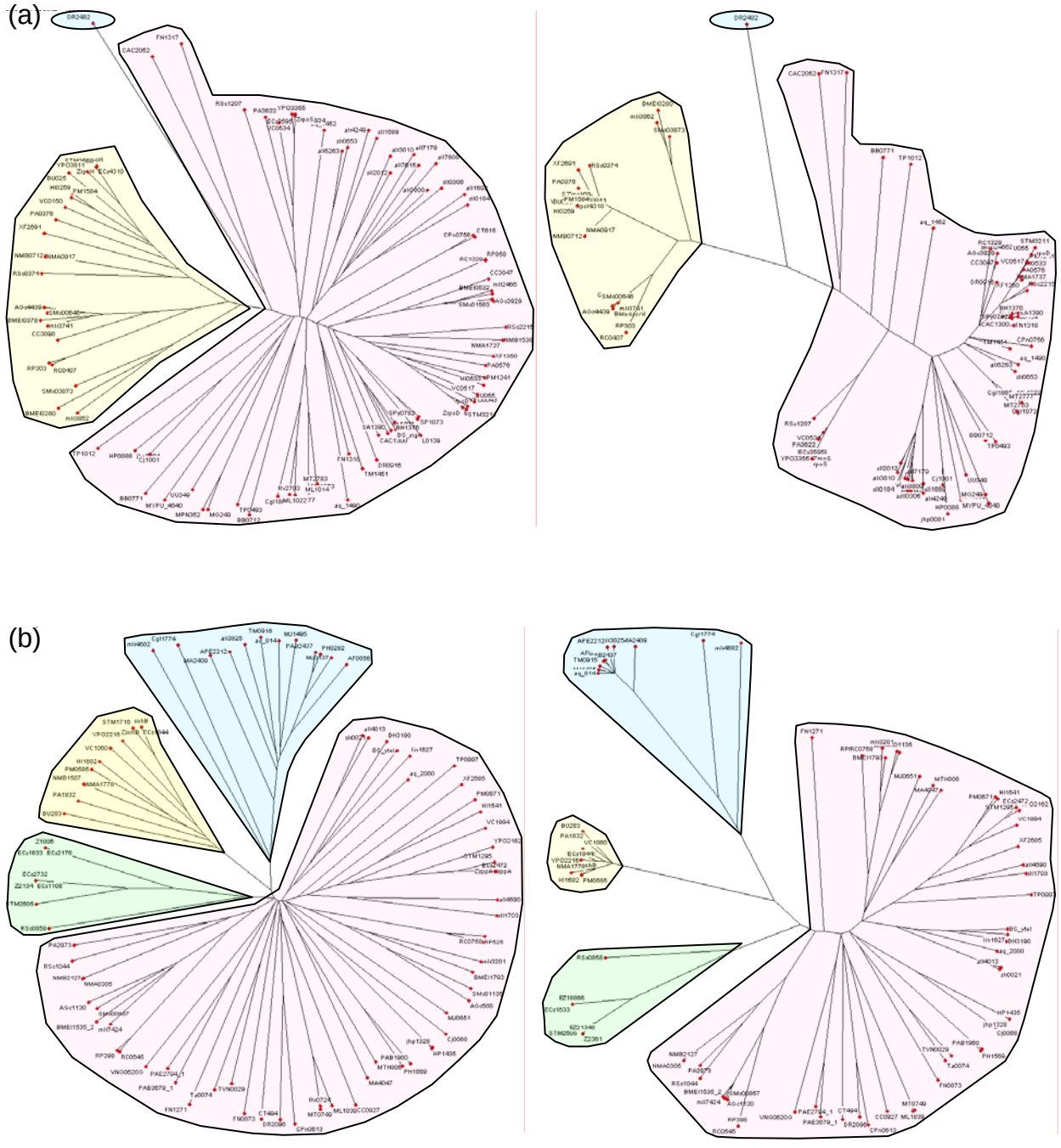


Fig 2: (a) BLAST best-hits (BeTs) graph used by Tatusov *et al*. (2003) to construct KOG3752 (ribonuclease H). Nodes with the same color are paralogs. Dark edges are mutual BeTs, and colored dotted edges are one-way BeTs. (b) Species tree for organisms in KOG_3752 (not drawn to scale). (c) Neighbor joining phylogenetic tree of KOG_3752 proteins (with edge bootstrap values) obtained from ClustalW alignment. (d) Neighbor joining phylogenetic tree of KOG_3752 proteins (with edge bootstrap values) obtained from T-Coffee alignment. Genes CE01984 (RNase H1), CE29157 (RNase H1A), and CE19682 (RNase 1C) correspond to the ones that appear in Arudchandran *et al*.'s (2002) analysis.

Tatusov, R. *et al*. (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, 4(1): 41.

Arudchandran, A. *et al*. (2002) Multiple ribonuclease H-encoding genes in the Caenorhabditis elegans genome contrasts with the two typical ribonuclease H-encoding genes in the human genome. *Mol. Biol. Evol.*, 19(11):1910-1919.


(a)
(b)

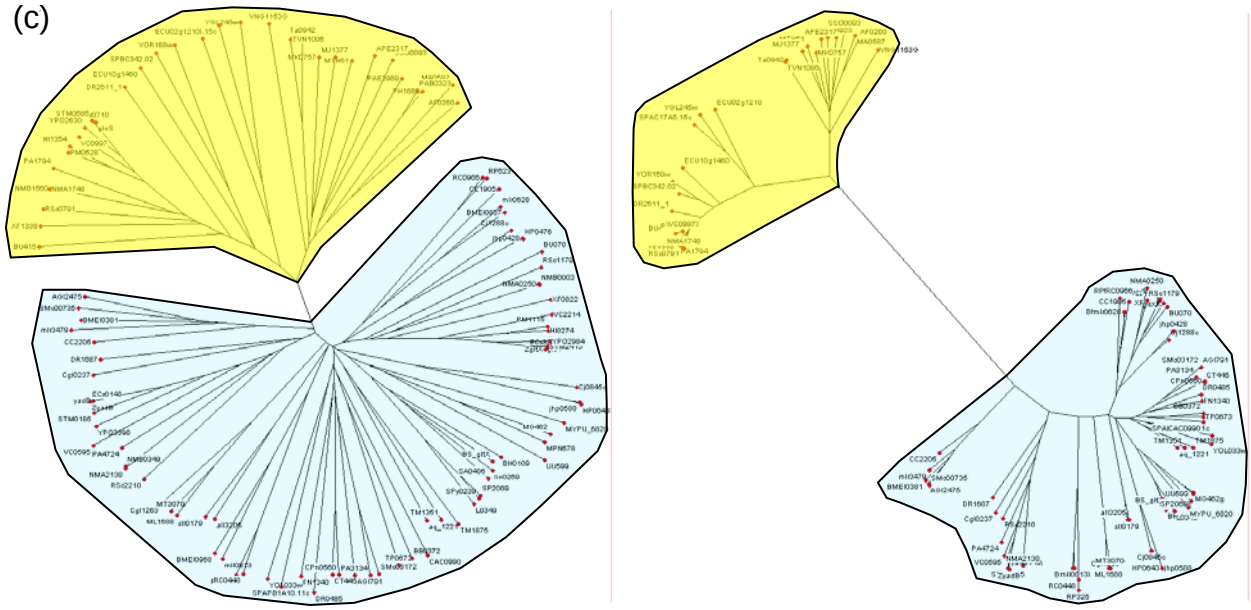


Fig 3: Neighbor joining phylogenetic trees constructed from the evolutionary distance matrix (on the left) and the correlation coefficient matrix (on the right). The clustering signal is clearly amplified when evolutionary correlations are used over evolutionary distances. (a) COG0568 (b) COG0616 (c) COG0008.

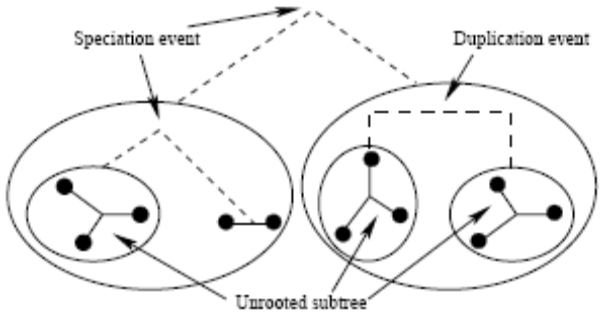


Fig 4: An illustrative example of a hierarchical clustering tree.