A Connected Linkage and Alignment Redefinition of COGs (CLARC) - clustering algorithm

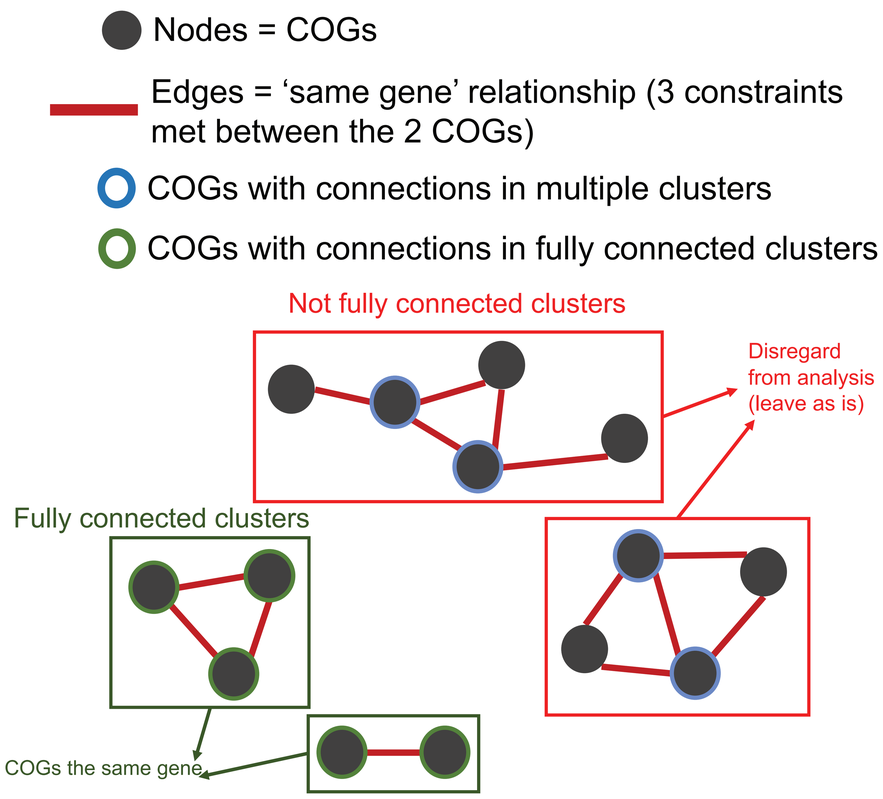
Pairwise identification of accessory COGs that look to be the 'same unit' based on constraints → Build graph and identify fully connected clusters → Re-define COGs; condense COGs in connected clusters

Two COGs called the 'same unit' if:

(1) COGs never co-occur in the same genome

(2) COGs have a high percentage of sequence similarity (default >95% identical matches BLASTN)

(3) COGs get classified in the same EggNOG functional group



	COG1	COG2
Sample 1	0	0
Sample 2	1	0
Sample 3	0	1
Sample 4	0	1

→

COG1-COG2
0
1
1
1

B General tool workflow