More info to pathways: the enrichment of COG and Kegg Orthology databases and the grouping of cognate proteins with Seed Linkage

J. Miguel Ortega

Lab. Biodados, Depto. Bioquímica e Imunologia, ICB, UFMG, Belo Horizonte, MG, Brasil

A demand for information on classified proteins is even larger after the emergence of a new generation of sequencing machines. At Laboratory of Biodados are providing an automated enrichment of an edited version of COG database and Kegg Orthology (KO) database as well. The procedure is based on the recruitment of additional members of UniRef50 clusters by their integrants that are present in the original databases followed by at least two filters: (i) size selection, where the recruited doe not diverge from recruiter from over 10% of recruiters size; (ii) taxonomic cut, where the recruited must belong to the same given clade (e.g. order) as the recruiter. Another complementary approach to form groups of related sequences is conducted by Seed Linkage software, that is able to cluster cognate proteins from multiple organisms beginning with only one sequence, through connectivity saturation with that seed sequence. Seed Linkage and UniRef enrichment are two procedures to enlarge and propagate information on homologous proteins, allowing the grouping of related amino acid sequences, that can be further used to investigate newly generated sequences.