Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

A procedure to recruit members to enlarge protein family databases - the building of UECOG (UniRef-Enriched COG Database) as a model

G.R. Fernandes[1]*, D.V.C. Barbosa[1]*, F. Prosdocimi[1], I.A. Pena[1], L. Santana-Santos[1], O. Coelho Junior[1], A. Barbosa-Silva[1], H.M. Velloso[1], M.A. Mudado[2], D.A. Natale[3], A.C. Faria-Campos[4], S.V. A. Campos[4] and J.M. Ortega[1]

[1]Departamento de Bioquímica e Imunologia, Laboratório de Biodados, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brasil
[2]Fundação Ezequiel Dias, Belo Horizonte, MG, Brasil
[3]Protein Information Resource, Georgetown University Medical Center, Washington, DC, USA
[4]Departamento de Ciência da Computação, ICEX, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brasil
*These authors contributed equally to this study.

Corresponding author: J.M. Ortega
E-mail: miguel@icb.ufmg.br

Genet. Mol. Res. 7 (3): 910-924 (2008)
Received June 2, 2008
Accepted August 11, 2008
Published September 30, 2008

ABSTRACT. A procedure to recruit members to enlarge protein family databases is described here. The procedure makes use of UniRef50 clusters produced by UniProt. Current family entries are used to recruit additional members based on the UniRef50 clusters to which they belong. Only those additional UniRef50 members that are not fragments and whose length is within a restricted range relative to the original entry are recruited. The enriched dataset is then limited to contain only genomes from selected clades. We used the COG database - used for genome annotation and for studies of phylogenetics and gene evolution - as a model. To validate the method, a

UniRef-Enriched COG0151 (UECOG) was tested with distinct procedures to compare recruited members with the recruiters: PSI-BLAST, secondary structure overlap (SOV), Seed Linkage, COGnitor, shared domain content, and neighbor-joining single-linkage, and observed that the former four agree in their validations. Presently, the UniRef50-based recruitment procedure enriches the COG database for Archaea, Bacteria and its subgroups Actinobacteria, Firmicutes, Proteobacteria, and other bacteria by 2.2-, 8.0-, 7.0-, 8.8-, 8.7-, and 4.2-fold, respectively, in terms of sequences, and also considerably increased the number of species.

Key words: COG; Secondary database; UniRef; UniProt; UECOG

