Reverse annotation and automated mining by using relational databases of proteomes from Model Organisms

[1]Alessandra C Faria-Campos, [1]Maurício A Mudado, [1]Adriano Barbosa-Silva, [2]João Torres, [1]Saulo Paula-Pinto, [1]Daniela VC Barbosa, [2]Sérgio Campos, [1]Glória R Franco, [1]J Miguel Ortega

[1]Depto. Bioquímica e Imunologia, [2]Depto. Computação Científica, Universidade Federal de Minas Gerais, 31270-010, MG, BRAZIL.

Gene annotation and mining usually starts when a novel transcriptome or genome project ends. Sequences are submitted to similarity searches using primary databases such as GenBank and the identification extracted from the best hits are attributed to the query sequence as the annotation string. After that, biologists search the annotated collection of sequences for proteins of interest, and further characterize it. We have devoted a great amount of effort on the proposal of reverse annotation and automated mining by using relational databases of proteomes from Model Organisms (MO): *A. thaliana*, *C. elegans*, *D. melanogaster* and *H. sapiens* . We have shown that sequences from MO both quantitatively and qualitatively may suffice for most of the process, assayed the performance of secondary databases such as COG/KOG, BioCarta, GOA and KEGG on selection of complete CDS from EST collections, tested the accuracy of automated mining with KOG proteins, developed a web toll that queries secondary databases and verifies domain structure and neighbor joining distances and provided a site with the EST sampling/expression of KOG proteins. Thus, our expertise supports the reverse annotation and automated mining using a previously populated secondary database consisting of a curate collection of proteins from Model Organisms. Protein Classification Tool (PCT) and KOG Expression/Sampling Tool (K-EST) are accessible at http://biodados.icb.ufmg.br.