



# Testing the performance of automated annotation of ESTs with the Kegg Orthology (KO) database demonstrates lack of completeness of clusters

G.R. Fernandes, M.A. Mudado and J.M. Ortega

Laboratório de Biodados, Departamento de Bioquímica e Imunologia, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brasil

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

Genet. Mol. Res. 7 (3): 948-957 (2008)  
Received June 2, 2008  
Accepted August 11, 2008  
Published September 30, 2008

**ABSTRACT.** The KEGG Orthology (KO) database was tested as a source for automated annotation of expressed sequence tags (ESTs). We used a control experiment where every EST was assigned to its cognate protein, and an annotation experiment where the ESTs were annotated by proteins from other organisms. Analyzing the results, we could assign classes to the annotation: correct, changed and speculated. The correct annotation ranged from 57 (*Caenorhabditis elegans*) to 81% (*Homo sapiens*). In spite of the changed annotation being low (1 in *H. sapiens* to 9% in *Arabidopsis thaliana*), the speculation was very high (18 in *H. sapiens* to 38% in *C. elegans*). We propose eliminating part of the speculated annotation using the KEGG Genes database to enrich KO clusters, decreasing the speculation from 38 to 2% in *C. elegans*. Thus, the KO database still demands some effort for moving sequences from Kegg GENES to KO, to complement the annotation performance.

**Key words:** KEGG Orthology; Annotation; Orthologs; BLAST; Expressed sequence tags