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

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

INTRODUCTION

Ever since the first complete genome of a cellular life form was described in 1995, the analyses and identification of genes and their function have become a challenge (Fleischmann et al., 1995; Brosius, 1996). Using computers to analyze the information from a well-understood organism, it is possible to transfer this knowledge to poorly characterized genomes, based on the similarity shared by their DNA or protein sequences (Tatusov et al., 1996). These similarity searches are usually conducted with the usage of several softwares available in the BLAST package (http://www.ncbi.nlm.nih.gov/BLAST/) (Altschul et al., 1997). A relationship of homology can be established by selecting the higher bit score and using an E-value cutoff requirement (Koonin and Galperin, 2003; Koonin et al., 2004). The quality of this type of annotation depends considerably on the quality, reliability and completeness of the information stored in the target database for BLAST alignment (Mudado et al., 2005). Secondary databases have been used as a source of information to annotate novel genome and transcriptome projects (Vettore et al., 2003). These databases usually contain genomics and proteomics data; furthermore, they provide function, structure, and other complementary information.

One of the best known secondary databases is KEGG (Kyoto Encyclopedia of Genes and Genomes (http://www.genome.jp/kegg). KEGG is a web resource directed at integrating genomic and functional information, and has been used to standardize gene annotation (Kanehisa and Goto, 2000). Up to November 2006, the KEGG database was composed of 1,775,476 gene entries derived from the genomes of 446 organisms (including 35 eukaryotes, 382 bacteria and 29 archea). The KEGG database information is organized into four sub-databases: GENES, BRITE, PATHWAY and LIGAND. GENES contains gene catalogs of completely sequenced genomes and some partial genomes. BRITE informs about protein-protein interactions and relationships. The PATHWAY database informs about generalized protein interaction networks (pathway and complexes), where several cellular processes are involved. LIGAND informs about chemical compounds and chemical reactions that are relevant to cellular processes (Kanehisa et al., 2002). KEGG Orthology (KO) was developed to integrate pathway and genomic information in KEGG. KO was introduced to replace the Enzyme Commission (EC) number as identifier of gene product in metabolic pathways. In order to classify gene functions, KO is based on computational analyses and manual curation of the SSDB (Sequence Similarity Database) ortholog clusters. KO is structured as a hierarchy of four flat levels, and this structure allows KO to be a great putative source for automated annotation (Mao et al., 2005).

In the present study, we tested the potential of the KO database to automatically annotate ESTs of model organisms using BLAST, with a methodology we proposed earlier (Mudado et al., 2005). Although the annotation of expressed sequence tags (EST) with KO is a common procedure, the tests of performance conducted here are informative and aimed at addressing the confidence of the method. We first assigned the ESTs of a given organism to their proteins deposited in KO. Afterwards, we used the remaining proteins from the other organisms in KO, without the proteins of the cognate organism used in the assignment, to annotate the previously assigned ESTs, evaluating the correctness of the results. The resulting data show that the annotation accuracy was considerably high. However, in spite of the changed annotation being low (1 in *Homus sapiens* to 9% in *Arabidopsis thaliana*), the speculated annotation (defined here as annotation in absence of assignment) was unexpectedly very high (18 in *H. sapiens* to 38% in *Caenorhabditis elegans*) as compared to results previously obtained with the KOG database (Mudado et al., 2005). We

showed evidence that many ESTs that receive speculated annotation can be assigned to protein entries in KEGG GENES. We propose eliminating part of the speculated annotation using the KEGG GENES database to enrich KO clusters, which markedly decreases 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.

MATERIAL AND METHODS

Download and selection of ESTs

ESTs were retrieved from the NCBI web site (www://ncbi.nlm.nih.gov). We downloaded collections of ESTs from four model organisms: *A. thaliana*, *C. elegans*, *Drosophila melanogaster*, and *H. sapiens*. We chose to download GenBank flat files and filter these files with a Perl script to select ESTs regarding information about, for example, organ, tissue, library, author, development stage, and sequence length. These data were used to populate an MySQL database, which was used to retrieve ESTs from healthy tissues and from libraries with more than five thousand entries (see Table 1).

Proteins

Protein sequences were downloaded from the KEGG web site (ftp://ftp.genome.jp/pub/kegg/). We used sequences from ten organisms in the analysis: *A. thaliana* (Ath), *C. elegans* (Cel), *Canis familiaris* (Cfa), *D. melanogaster* (Dme), *Encephalitozoon cuniculi* (Ecu), *H. sapiens* (Hsa), *Mus musculus* (Mmu), *Rattus norvegicus* (Rno), *Schizosaccharomyces pombe* (Spo), and *Saccharomyces cerevisiae* (Sce). Only proteins present in the KO database were used, which represent a small proportion of the total available proteins from KEGG GENES, as shown in Figure 1.

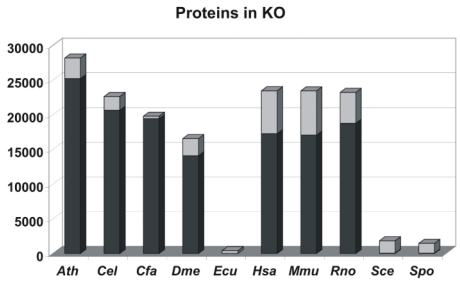

Figure 1. Distribution of the organism's proteins into KEGG Orthology (KO) clusters. The light grey bars represent the proteins within KO, while dark grey bars those outside KO.

BLAST experiments

The BLAST experiment was divided into two stages (Figure 2): initially, we executed a tBLASTn using the KO proteins from one organism against its own EST repository (e.g., Ath

proteins *vs* Ath ESTs). This stage works as a positive control, which assigns a given EST to its cognate protein in the database. The low complexity filter was disabled and an E-value cutoff of at least 10^{-10} was established in order to favor alignments of only homologous sequences. Additionally, an assignment identity cutoff was applied only for this step. Assignment cutoffs were determined as described by Mudado et al. (2005 and Mudado MA, Fernandes GR and Ortega JM, unpublished results). In the second step, we aligned the ESTs with the database depleted of the cognate organism proteins. We used the same threshold for low complexity filter and E-value established above. This experiment was aimed at simulating the annotation procedure of a novel transcriptome, as the ESTs were aligned only with proteins from organisms different from its source organism. A summary of both stages is shown in Figure 2. This procedure has already been performed with the KOG database (Mudado et al., 2005) in order to evaluate the database potential of annotating.

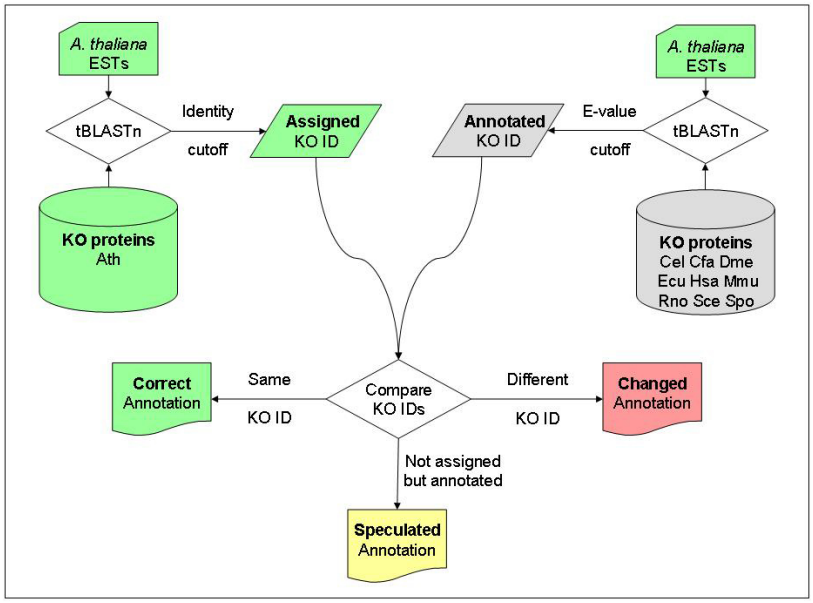

Figure 2. Representation of the performance tests using expressed sequence tags (ESTs) from *Arabidopsis thaliana*. On the left, alignments requiring an identity cutoff (78% for *A. thaliana*) on top of a 10^{-10} E-value cutoff (data not shown) serve as a positive control, and assignment of a KEGG Orthology (KO) identifier (KO ID) to each EST. On the right, a second step simulates a novel genome annotation. Comparing the results, we labeled the annotation as correct, changed or speculated.

Assignment cutoff

Briefly (see Mudado et al., 2005, for details), 50 proteins without paralogs with the highest expression determined by the number of hits to EST were selected for analysis. ESTs were aligned to either the nucleotide sequence of those proteins' CDS with BLASTn or to those proteins with tBLASTn, under a stringent E-value cutoff of 1e-10 (note that to preserve the E-value determination the formatted database used in both steps was the total set of ESTs).

We first determined the percentage of aligned ESTs that show over 96% identity at the nucleotide to nucleotide level. We then determined the identity cutoff at the nucleotide to amino acid level, which groups this same percentage of hits, and used this value as identity cutoff for EST assignment to the correct protein.

Annotation classes

Classes of annotation were established by comparing the BLAST results of the stages 1 (assignment) and 2 (annotation). ESTs within the correct annotation category were assigned and annotated by the same KO protein from stages 1 and 2. Changed annotation category has ESTs with different assigned and annotated KO proteins. Speculated annotation occurs when an EST is annotated by a KO protein but it has not been assigned by any protein (so we have no positive control for that annotation).

KEGG GENES

The KEGG GENES database was also used to assign ESTs in order to try to minimize speculated annotation. This database involves all proteins from an organism that is present in KEGG.

Accuracy

The accuracy was measured by the PPV (positive predictive value). In this case the PPV can be defined as the quotient of the correct annotation divided by the correct and changed annotations.

RESULTS AND DISCUSSION

Annotation overview

After downloading and filtering all ESTs (Table 1), the BLAST searches were performed and the annotation divided into categories (see Material and Methods).

Table 1. Total number of expressed sequence tag (EST) downloaded and total remaining after MySQL filtering.

Organism	EST - Downloaded	EST - Used
Arabidopsis thaliana	622,788	360,833
Caenorhabditis elegans	302,080	293,530
Drosophila melanogaster	383,407	375,360
Homo sapiens	1,673,145	365,619

With the default identity cutoffs (78, 81, 71, and 72%), annotation was distributed as follows: 71, 57, 70, and 81% of correct annotation; 9, 5, 5, and 1% of changed annotation, and 20, 38, 25, and 18% of speculated annotation, for Ath, Cel, Dme, and Hsa, respectively. Although changed annotation was low, the database provided very high levels of speculated annotation, suggesting lack of completeness of the KO clusters.

The total annotation covered 21, 36, 32.5, and 37.1% of all ESTs from Ath, Cel, Dme, and Hsa, respectively. By imposing an identity cutoff, the proportion of correct annotations decreased and the proportion of speculated annotation increased until it reaches a less variable level if the assignment cutoff is between 70-80% identity (Figure 3; see also discussion about accuracy below). This effect is mostly caused by a decrease in EST assignment if the cutoff is too stringent in the first stage of the procedure (Figure 2). However, the changed annotation does not show apparent changes as the cutoff varies. Based on previous results with the KOG database, we believe that changed annotation is mainly caused by proteins that are specific to an organism but can be annotated by similar proteins, which share the same conserved domains. With the KOG database (Mudado et al., 2005), a clear plateau is observed with low values of assignment cutoffs, and that is possibly explained by the more completeness of this database as compared to KO (see below). Although the cutoff experimentally defined by us has experimental support, the results were obtained with all possible cutoffs for comparison, and the conclusion is that the correct and changed annotations both do not vary so much if a lower cutoff is chosen.

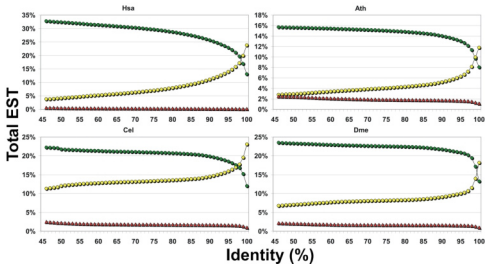

Figure 3. Proportion of the annotation classes and their relationship with the identity cutoff. The green circles represent the correct annotation, the yellow circles are the speculative annotation, and the red triangles represent the changed annotation. EST = expressed sequence tag.

Coverage

The KO coverage (percentage of KO clusters that find a hit in a representative collection of ESTs) is an important aspect to be analyzed, since it tells how relevant the KO groups are for the annotation of reasonably large collections of ESTs. The coverage overview is shown in Figure 4. KO includes several organisms and a large set was used in this study (Figure 1). Taking into consideration the KO groups of all these organisms (upper panels), it was observed that the coverage during the annotation (referred to here as "Sampling" of the EST collection with KO clusters, panels on the left) was higher than during the assignment (referred to here as "Picturing" the EST content supposedly assigned to the correct protein, panels on the right). Hsa clusters apparently are the most complete. Accordingly, this effect is reduced when only KO clusters that contain proteins of the studied organism are used (lower panels). Here, Sampling yields almost the same coverage as Picturing (even for Ath, the only plant), indicating that similarity between cluster members is sufficient. Thus, coverage does not seem to be substantially affected by the lack of paralogs in clusters that contain at least a representative of the organism whose ESTs are being analyzed.

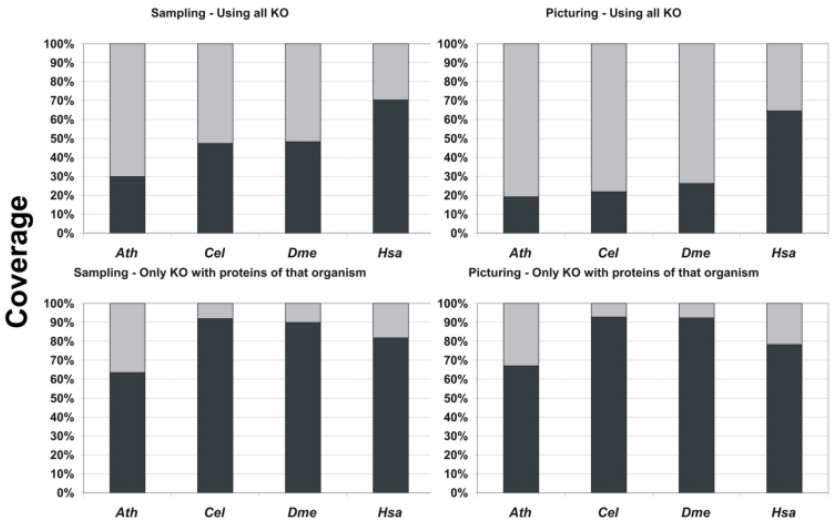

Figure 4. Percentage of KEGG Orthology (KO) clusters used for the expressed sequence tags (EST) assignment (picturing) and annotation (sampling) of Ath, Cel, Dme, and Hsa. In the two upper graphics, we used the KO clusters of all available organisms, while in the graphics at the bottom, we used as reference only the KO groups, which contain proteins of the same organism to which the EST belongs. The dark bars indicate the KO groups that had a protein that matched with an EST, and the light bars represent the cluster that had no matches with this EST collection.

Extra-KO annotation

The KEGG GENES database was complementarily used in order to try to minimize speculated annotation. This database contains all identified proteins of an organism present in KEGG. A graphical representation of this search is shown in Figure 5.

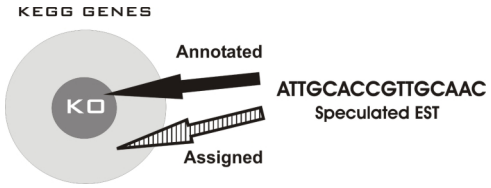

Figure 5. Procedure of the extra-KO annotation. An expressed sequence tag (EST) that had a speculated annotation was submitted to a BLAST against all KEGG GENES proteins of the same EST's organism. We followed the same parameters of E-value cutoff and low complexity filter used in the previous BLAST experiment. The same identity cutoff was applied to determine the assignment. KO = KEGG Orthology.

Some examples of KO clusters related with speculated annotation, and their respective extra-KO matches are shown in Table 2. The first column shows the organism to which the EST belongs. The second column represents the cluster entry that contains the protein, which matched the EST leading to the speculated annotation. The third shows how many times that KO cluster was related to a speculation. The fourth and fifth columns show the KO cluster name and

the extra-KO protein name, respectively. The data suggest that speculated annotation could turn into correct annotation as the KO clusters are enlarged to group entries from KEGG GENES.

Table 2. Relationship of speculation and extra-KO assignment.

Organism	KO	Amount	Cluster name	Extra-KO protein name
Ath	K05692	805	Actin, beta/gamma, cytoplasmic	F27J15.1; actin 8 (ACT8)
Cel	K01829	2622	Protein disulfide-isomerase	Protein disulfide isomerase protein 2, isoform a
Dme	K04439	1699	Arrestin	Arrestin

KO = KEGG Orthology.

Considering the cutoff values of 78, 81, 71, and 72% for Ath, Cel, Dme, and Hsa, respectively, the distribution of the annotation classes after the "extra-KO" experiment is shown in Figure 6. Note that the light green area would consist of speculated annotation (yellow) if the KEGG GENES assignment was not considered. This area is highlighted in light green in supposition that its candidates can be turned into correct annotation.

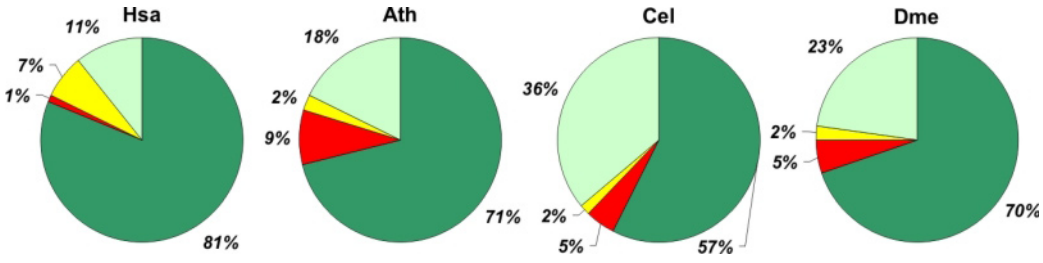

Figure 6. Pie graphs indicating the distribution of the annotation classes with an identity cutoff value of 72, 78, 81, and 71% for Hsa, Ath, Cel, and Dme, respectively. The dark green area represents the correct annotation, the light green represents the EST with speculated annotation, which has matches in the extra-KO proteins, the yellow area represents the speculated annotation that was not solved by the extra-KO assignment, and the red area indicates the changed annotation proportion.

Since *C. elegans* was the organism demonstrating the highest proportion of speculated annotation, we decided to evaluate a group of biochemical pathways involved in amino acid metabolism to illustrate this point. Table 3 lists sixteen pathways as obtained from the KO database and the number of KO entries in them, adding up to 692 KO clusters. Note that some pathways either consist of or include non-essential amino acid metabolism, and thus, *C. elegans* may lack proteins in some clusters such as methionine metabolism. However, we detected speculated annotation of ESTs to those clusters in 186 events. In a total of 152 cases, we were able to assign those ESTs to proteins from the KEGG GENES database that were not included in KO, indicating that these proteins are prompted to be included in the KO clusters. Moreover, when sequences from certain KO pathways are selected to evaluate expression in other organisms such as worms, the analysis could be compromised by the lack of entries in the KO database. We believe that the high levels of speculated annotation observed are probably due to a cautious but incomplete evolution of KO clusters as compared to the KOG database.

Table 3. Number of KEGG Orthology (KO) clusters from amino acid metabolism pathways that yield speculated annotation of *Caenorhabditis elegans* expressed sequence tags (ESTs) but with ESTs assigned to Kegg GENES extra-KO *C. elegans* proteins.

Pathway*	Number of KO in pathway	Lacking cel proteins	With speculated annotation	With extra-KO assignment
Total	692	511	186	152 (22%)

*Consist of or include biosynthesis pathway of non-essential amino acids.

Accuracy

The annotation method accuracy for each organism is shown in Figure 7, where the curves represent the PPV value fluctuation for each organism. The best accuracy was obtained for the Hsa annotation. It can be explained by the fact that the KO database uses this organism as reference for the ortholog groups. The worst accuracy, but acceptable, was obtained for Ath. This might have occurred because it is the only plant in the database.

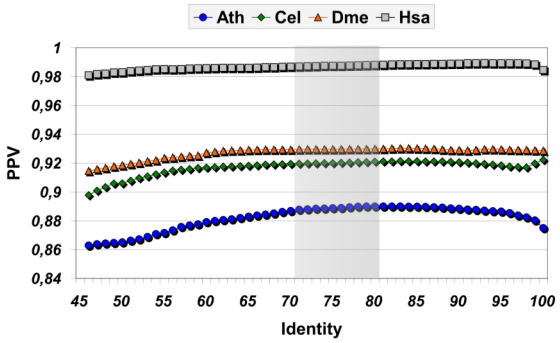

Figure 7. Annotation accuracy fluctuation related to identity among the sequences. Note: identity in percentages, and positive predictive value (PPV) varying from 0-1.

CONCLUSION

We demonstrated that, although being an incomplete database, KO is a good vocabulary source for automated annotation. This was proved by the high accuracy when assigned and annotated datasets were compared among the distinct organisms that build KO. For this purpose, it is also crucial that identity cutoff be selected in such a way that a great coverage of correctly annotated ESTs balances with a low proportion of changed annotation. The PPV test gave us an overview of this value, which allowed us to confirm the appropriated cutoff values. Additionally, the lack of entries that are present in KEGG GENES can explain the high amount of speculated annotation, suggesting that additional evolution of the KO database with inclusion of entries in KEGG GENES will greatly enhance the performance of automated annotation with KO. This could be seen in the extra-KO experiment, where proteins without a KO cluster were used to minimize the speculation. In summary, automated annotation with KO is accurate, and the analysis presented here supports the inclusion of KEGG GENES entries in KO clusters for enhanced performance of annotation with the KO database.

ACKNOWLEDGMENTS

We thank Dr. Darren Natale from PIR for critically reviewing this manuscript. G.R. Fernandes and M.A. Mudado were recipients of fellowships from CAPES. Laboratório de Biodados had a grant from FAPEMIG as a member of Minas Gerais Genome Network.

REFERENCES

Altschul SF, Madden TL, Schaffer AA, Zhang J, et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389-3402.

Brosius J (1996). More *Haemophilus* and *Mycoplasma* genes. *Science* 271: 1302-1304.

Fleischmann RD, Adams MD, White O, Clayton RA, et al. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269: 496-512.

Kanehisa M and Goto S (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28: 27-30.

Kanehisa M, Goto S, Kawashima S and Nakaya A (2002). The KEGG databases at GenomeNet. *Nucleic Acids Res.* 30: 42-46.

Koonin EV and Galperin MY (2003). Sequence-Evolution-Function. Computational Approaches in Comparative Genomics. Kluwer Academic Publishers, Norwell.

Koonin EV, Fedorova ND, Jackson JD, Jacobs AR, et al. (2004). A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol.* 5: R7.

Mao X, Cai T, Olyarchuk JG and Wei L (2005). Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. *Bioinformatics* 21: 3787-3793.

Mudado MA, Bravo-Neto E and Ortega JM (2005). Tests of automatic annotation using KOG proteins and ESTs from 4 eukaryotic organisms. *Lecture Notes Computer Sci.* 3594: 141-152.

Tatusov RL, Mushegian AR, Bork P, Brown NP, et al. (1996). Metabolism and evolution of *Haemophilus influenzae* deduced from a whole-genome comparison with *Escherichia coli*. *Curr. Biol.* 6: 279-291.

Vettore AL, da Silva FR, Kemper EL, Souza GM, et al. (2003). Analysis and functional annotation of an expressed sequence tag collection for tropical crop sugarcane. *Genome Res.* 13: 2725-2735.

