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

A picture of gene sampling/expression in model organisms using ESTs and KOG proteins

Maurício de Alvarenga Mudado and José Miguel Ortega

Laboratório de Biodados, Departamento de Bioquímica e Imunologia,
Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais,
Av. Antônio Carlos, 6627, Pampulha, Caixa Postal 486,
31270-010 Belo Horizonte, MG, Brasil
Corresponding author: J.M. Ortega
E-mail: miguel@ufmg.br

Genet. Mol. Res. 5 (1): 242-253 (2006)
Received January 10, 2006
Accepted February 17, 2006
Published March 31, 2006

ABSTRACT. The expressed sequence tag (EST) is an instrument of gene discovery. When available in large numbers, ESTs may be used to estimate gene expression. We analyzed gene expression by EST sampling, using the KOG database, which includes 24,154 proteins from *Arabidopsis thaliana* (Ath), 17,101 from *Caenorhabditis elegans* (Cel), 10,517 from *Drosophila melanogaster* (Dme), and 26,324 from *Homo sapiens* (Hsa), and 178,538 ESTs for Ath, 215,200 for Cel, 261,404 for Dme, and 1,941,556 for Hsa. BLAST similarity searches were performed to assign KOG annotation to all ESTs. We determined the amount of gene sampling or expression dedicated to each KOG functional category by each model organism. We found that the 25% most-expressed genes are frequently shared among these organisms. The KOG protein classification allowed the EST sampling calculation throughout the glycolysis pathway. We calculated the KOG cluster coverage and inferred that 50 to 80 K ESTs would efficiently cover 80-85% of the KOG database clusters in a transcriptome project. Since KOG is a database biased towards housekeeping genes, this is probably the number of ESTs needed to include the more commonly expressed genes in these organisms. We also examined a still unaddressed question: what is the minimum number of ESTs that should be produced in a transcriptome project?

Key words: EST, Transcriptome projects, KOG, COG, Annotation

INTRODUCTION

The expressed sequence tag (EST) is an instrument of gene discovery. Although it bears around 3-4% sequencing errors (Hillier et al., 1996), this tag suffices for identification of orthologous genes in other organisms through homology searches, thus providing functional annotation of the EST and also demonstrating the presence of the gene in the organism and/or developmental stage of interest (Adams et al., 1991; Faria-Campos et al., 2003). Large numbers of ESTs, coming from independent cDNA libraries, can be used to estimate gene expression (Lee et al., 1995; Franco et al., 1997; Ewing et al., 1999). EST occurrence also allows a gene sampling estimate in novel transcriptome projects.

The total number of EST sequences deposited in public databases has grown considerably (see http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html). Transcriptome projects are still a good alternative to genome projects as they are less expensive and generate information about gene expression (reviewed by Lindlof, 2003). Large numbers of EST sequences are being produced, although there are still some questions to be answered regarding this approach, such as: i) Is transcript redundancy really directly connected to gene expression? ii) How many cDNA libraries from different tissues and developmental stages are needed to yield a complete picture of an organism's transcriptome? iii) What is the minimum number of ESTs that need to be produced in a transcriptome project in order to give a good representation of the most ubiquitous genes (e.g., housekeeping genes)? We evaluated this last point.

An initial approach to answer this question is to perform automatic transcriptome annotation using secondary databases, where ESTs are annotated by similarity to characterized proteins. The KOG database (eukaryotic representatives of the COG database - Tatusov et al., 2001 and 2003) is one of the many secondary databases, such as GOA/UniProt and KEGG (Kanehisa and Goto, 2000; Camon et al., 2003) that have sequences classified into functional categories and groups, and can be used for this kind of study. The KOG database contains 24,154 proteins from *Arabidopsis thaliana* (Ath), 17,101 from *Caenorhabditis elegans* (Cel), 10,517 from *Drosophila melanogaster* (Dme), and 26,324 from *Homo sapiens* (Hsa). KOG proteins are clustered by function so it is plausible to assume that the KOG database is biased towards ubiquitous clusters - genes that are simultaneously present in at least three model organisms among the seven composing the database.

In order to estimate the efficiency of gene sampling in transcriptome projects, BLAST similarity searches were conducted using ESTs and KOG proteins from these four organisms, requiring different similarities for different organisms. We previously determined that similarity cutoffs should be 78% for Dme and Cel, 80% for Hsa and 84% for Ath (Mudado et al., 2005).

Briefly, experiments were conducted with either pUC18 sequence reads or ESTs; the alignment of these sequences to their respective edited nucleotide sequences was selected based on an identity cutoff of 96% (since single-pass reads may bear up to 4% errors); the similarity cutoff determined for their alignments to the respective edited amino acid sequences was over 80%.

The KOG database allows comparison of all the gene samples of the organisms by functional categories, or with respect to a specific pathway, such as glycolysis, since enzymes that compose the pathways have already been classified in the database. In addition it is possible to compare genes sampled simultaneously from one up to all organisms present in the database.

Public EST databases and secondary databases are depositories of novel and growing information that can be extracted with appropriate bioinformatics approaches. We made use of dbEST (Boguski et al., 1993) and KOG databases, from the NCBI, for sampling of KOG genes within the transcriptome collection available for four model organisms. This approach allows one to predict the number of EST sequences that need to be produced in order to represent the genes present in KOG. We also estimated the minimum number of reads necessary in a novel transcriptome project.

MATERIAL AND METHODS

BLAST

The EST sequences were downloaded from the dbEST database by May 2003. All KTL (KOG, TWOG and LSE) proteins and KOG-conserved domains were downloaded from NCBI (ftp://ftp.ncbi.nih.gov/pub/COG/KOG/) and were in the “kyva” file. All ESTs and proteins that were used are shown in Table 1.

BLAST software (version 2.2.8) was used to search EST sequences against a KOG database depleted of conserved domains. We selected only the 88,613 classified KTL proteins found in the “kog”, “twog” and “lse” files at the KOG site, for use in the BLAST searches as queries. The KTL proteins were divided into 60,758 KOG, 4,451 TWOG and 23,404 LSE proteins. The subjects of the BLAST searches were the organisms’ ESTs. tBLASTn was used with the following parameters: -m 8 -b 10e6 -e 1e-10 -F f. These parameters activate the tabular output of BLAST, allowing up to 10 million hits to one protein (default is 250) and deactivates the low-complexity filter, respectively. The low-complexity filter was deactivated in order to allow tBLASTn to achieve 100% identity in the alignments.

Data processing

All BLAST results, obtained in tabular output (m8 option) with an e-value cutoff of 10^{-10}, were added to an MySQL (version 3.23.58) database, which was populated with all the information related to the KOG database, including the functional classification assigned to each protein. When necessary, PERL (version 5.8.0) scripts were generated to solve computational problems.

The sampling/expression was defined by the number of ESTs that were a best hit in the BLAST searches and the number of hits per KOG functional category was counted. The best scores were always selected to avoid assigning any given EST to more than one protein. Similarity cutoffs of 78% for Cel and Dme, 80% for Hsa and 84% for Ath were used, as previously described (Mudado et al., 2005).

To create random EST datasets, PERL scripts were created with the Math::Random package (thanks to John Venier and Barry W. Brown) downloaded from CPAN (http://www.cpan.org). Ten thousand up to 150,000 ESTs were selected in increasing rounds of 10,000 EST selections. Every round was repeated 10 times in order to obtain the sampling error. All EST selections were made in an independent manner (in every round the ESTs were reselected from the database).

The KOG coverage was calculated in two different manners: cluster coverage (Figures 5 and 6) and protein coverage (Figure 7). A KOG cluster was assumed “covered” when at least one protein from that cluster had a hit to an EST sequence. On the other hand, protein coverage was stricter, as it demanded that all KOG proteins from one cluster had hits in order to yield 100% coverage. The KOG coverage was calculated using only KOG clusters that represent genes of at least three model organisms. TWOGs and LSEs were not included in incremental experiments (Figures 5 to 7), since they contain too many non-categorized proteins (functional category X), which are not well annotated (unnamed proteins) and are more organism-specific than KOGs (see supplementary Tables S2 and S3 at http://biodados.icb.ufmg.br). However, experiments including TWOGs and LSEs are shown in supplementary Figures S4 to S6. Supplementary material is available at Laboratório de Biodados, UFMG (http://biodados.icb.ufmg.br). To perform the cluster coverage and protein coverage experiments, 4,597, 3,285, 4,235, and 4,351 KOG clusters were selected, which contained 19,039, 13,744, 10,581, and 8,445 proteins from Hsa, Ath, Cel, and Dme, respectively.

Statistical analysis

Data were reported as means ± SEM (standard error of the mean). All *t*-tests performed were unpaired, with 50 degrees of freedom.

RESULTS AND DISCUSSION

Gene sampling

A set of ESTs corresponding to all ESTs available for Ath, Cel and Dme (by May 2003; Table 1) was downloaded. About 10 times more ESTs from Hsa were downloaded (almost 2 million), which was assumed to be a sufficiently large sample to obtain a precise analysis of Hsa gene sampling. Large sets of ESTs tend to dilute biases in cDNA libraries (e.g., more libraries from a specific organ or specific time of development) and redundant sequence deposits in dbEST. Conversely, EST production driven by sequencing centers is balanced to cover the transcriptome. However, the term ‘gene expression’ was avoided and substituted by the term ‘gene sampling’, as the main goal was not the EST per gene index, but the probability of gene discovery in a transcriptome project. All sampling data are available as part of K-EST: the KOG expression/sampling tool (http://biodados.icb.ufmg.br/K-EST/, Mudado et al., submitted).

Individual sampling lists were generated for each of the four organisms, resulting in sampling profiles by functional category (Figure 1). This methodology tends to depict, by the amount of gene sampling, how each organism differentially produces transcripts related to the

Table 1. Numbers of sequences used for comparing gene expression.

Organisms	ESTs	KTLs	KTL proteins	KOGs	KOG proteins
Arabidopsis thaliana	178,538	4,872	24,154	3,285	13,744
Caenorhabditis elegans	215,200	5,306	17,101	4,235	10,581
Drosophila melanogaster	261,404	5,145	10,517	4,351	8,445
Homo sapiens	1,941,556	6,572	26,324	4,597	19,039

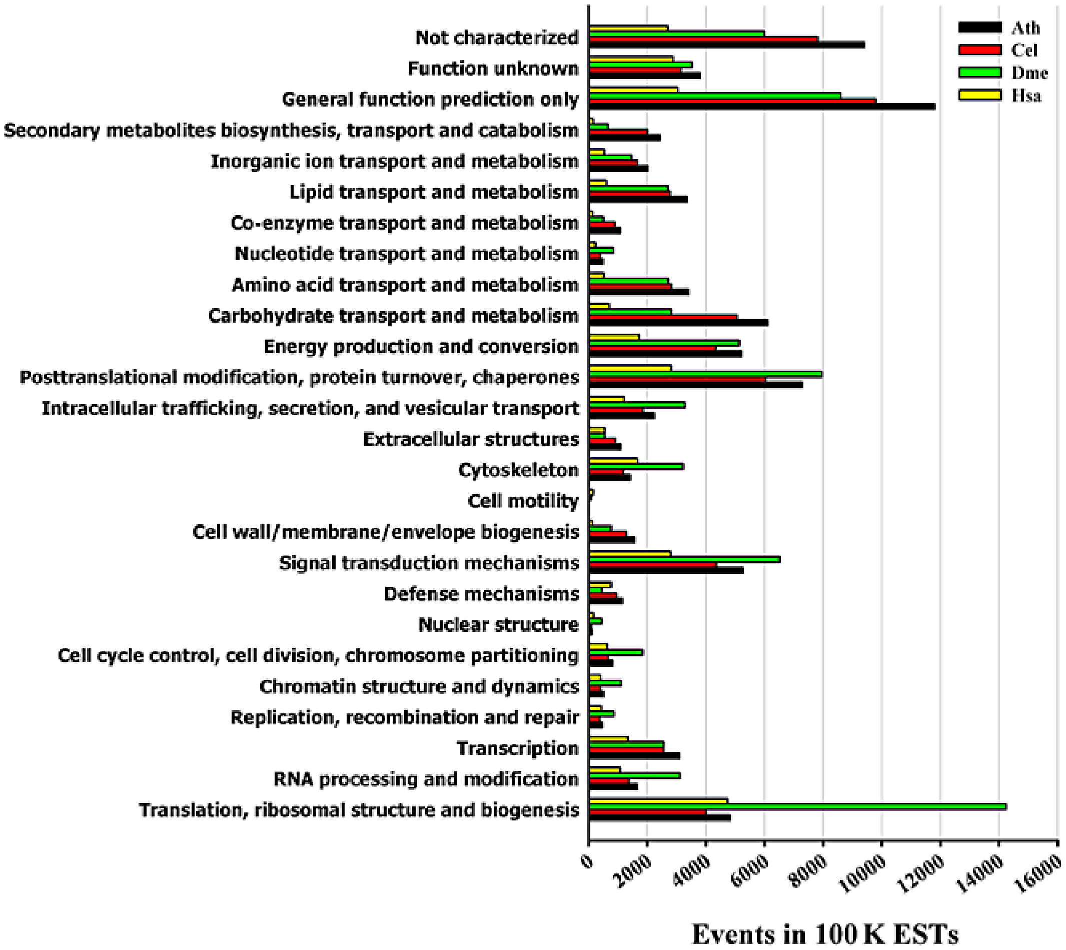

Figure 1. Gene sampling using KOG functional categories. The black, red, green, and yellow bars represent *Arabidopsis thaliana* (Ath), *Caenorhabditis elegans* (Cel), *Drosophila melanogaster* (Dme), and *Homo sapiens* (Hsa), respectively.

various types of biological processes. When doing the BLAST searches for gene sampling of the organisms, the other proteins were preserved in the query, as the best hits were almost totally composed of the cognate organism proteins. Most of the proteins of Cel, Ath and Hsa annotated their own ESTs (less than 0.5% of the annotation was from the proteins of other organisms). Only Dme appears to have had a more considerable cross-annotation, since around 3.8% of other organisms' KOG proteins were used to annotate its ESTs.

The differences in sampling of genes amongst organisms were analyzed. In Figure 2A, the 25% most and least expressed genes from the set of 2,523 KOG genes common to the four organisms were examined to determine the proportion of sharing involving most/least categories. As expected, all four eukaryotes shared 50-62% of the genes in the 25most category, while sharing 36-40% of the genes in the 25% least sampled set of genes, indicating that the more frequently sampled genes are shared more often amongst these organisms (P < 0.05). Evolutionary distance seems to count, since Dme and Cel share more genes per category than when they are compared with Ath and Hsa. The next step was to compare the set of KOG genes common to the four organisms, among all organisms for the two categories (25most and 25least). Figure 2B shows that sharing genes is more common in the 25most category. The 25least category produced the inverse situation, as expected.

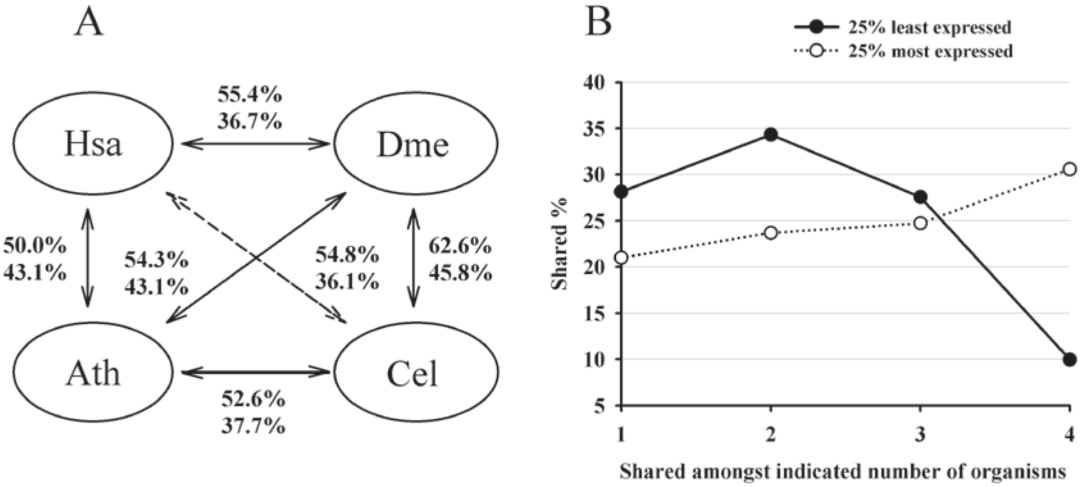

Figure 2. A. Comparison of the 25% most and least expressed genes for all four organisms. The upper and lower numbers of all tuples represent the 25% most and least expressed genes, respectively. **B.** Global comparison of the 25% most and least expressed genes among all four organisms. Hsa = *Homo sapiens*; Dme = *Drosophila melanogaster*; Ath = *Arabidopsis thaliana*; Cel = *Caenorhabditis elegans*.

The KOG database allows a more detailed analysis of gene sampling, since all genes are described and classified individually. Enzyme sampling within the glycolysis pathway was examined as an example. Figure 3 illustrates the similarities in the pathway of the four organisms. GAPDH was highly sampled in all four eukaryotes, followed by fructose biphosphate aldolase.

KOG coverage

The global cluster coverage was calculated for all functional categories in order to determine the intensity at which the KOG database clusters were covered by all organisms' ESTs. KOGs, TWOGs and LSEs specific for every organism were selected for this objective.

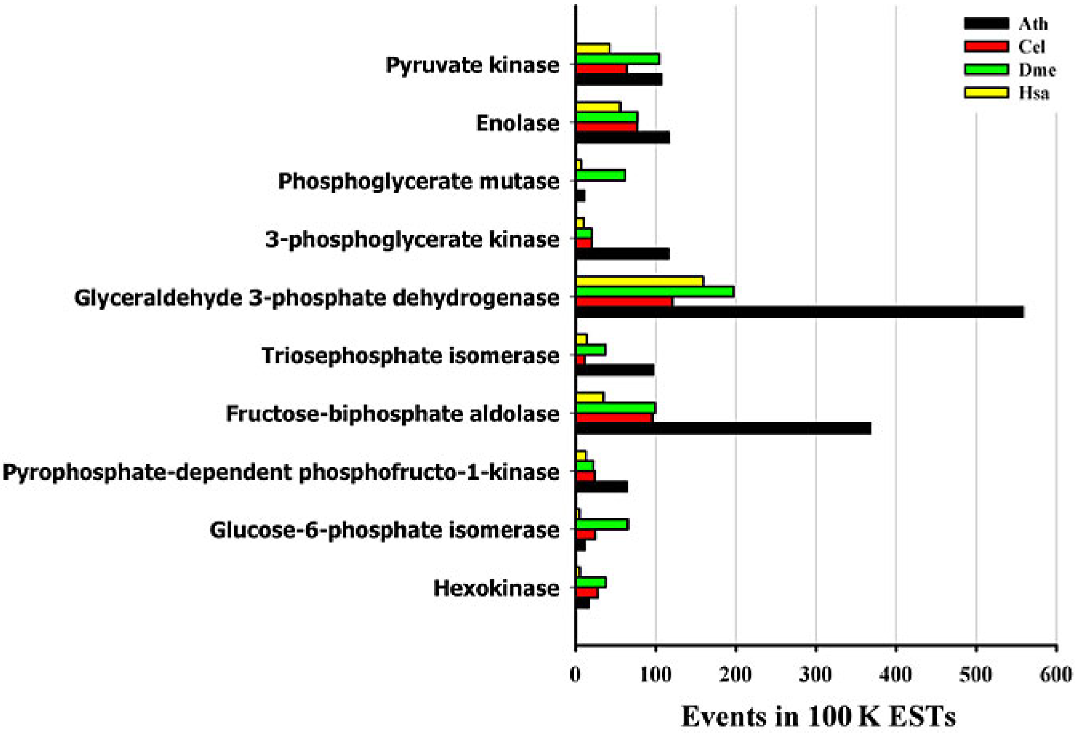

Figure 3. Sampling of glycolysis pathway enzymes. Ath = *Arabidopsis thaliana*; Cel = *Caenorhabditis elegans*; Dme = *Drosophila melanogaster*; Hsa = *Homo sapiens*.

The number of ESTs used appears to cover more than 95% of the KTL clusters of the four organisms (Figure 4).

In order to estimate an optimal number of ESTs that are sufficient to fully cover the KOG clusters, transcriptome projects with different sizes were simulated by creating EST pools, selected at random, from the EST databases of each organism. We selected only KOG entries (genes present in at least three model organisms) specific for each organism. TWOGs and LSEs were discarded since they represent genes that are more organism-specific. Two distinct experiments were executed. First, by maintaining the KOG proteins from all organisms in the database, curves of coverage tended to saturate (Figure 5A). Sets from 10 to 150 K ESTs were generated, with 10 repetitions in order to obtain the sampling errors. Saturation of the coverage curve was expected to occur since all ESTs probably have their correlated proteins in the database. A similar coverage result was expected when annotating novel ESTs, using databases that contain proteins with high similarity to the query organism (closely related organisms). The cluster coverage rose exponentially when using 10 to 80 K ESTs and then increased in a linear pattern. We suggest that 50 to 80 K is a reasonable minimum number of ESTs to be produced in order to obtain around 80-85% of the genes that are common among organisms, such as the genes represented by KOG. Also, all organisms but Hsa required at least 60 K ESTs to cover 80% of KOG clusters. *Homo sapiens* ESTs had a different behavior and showed less coverage

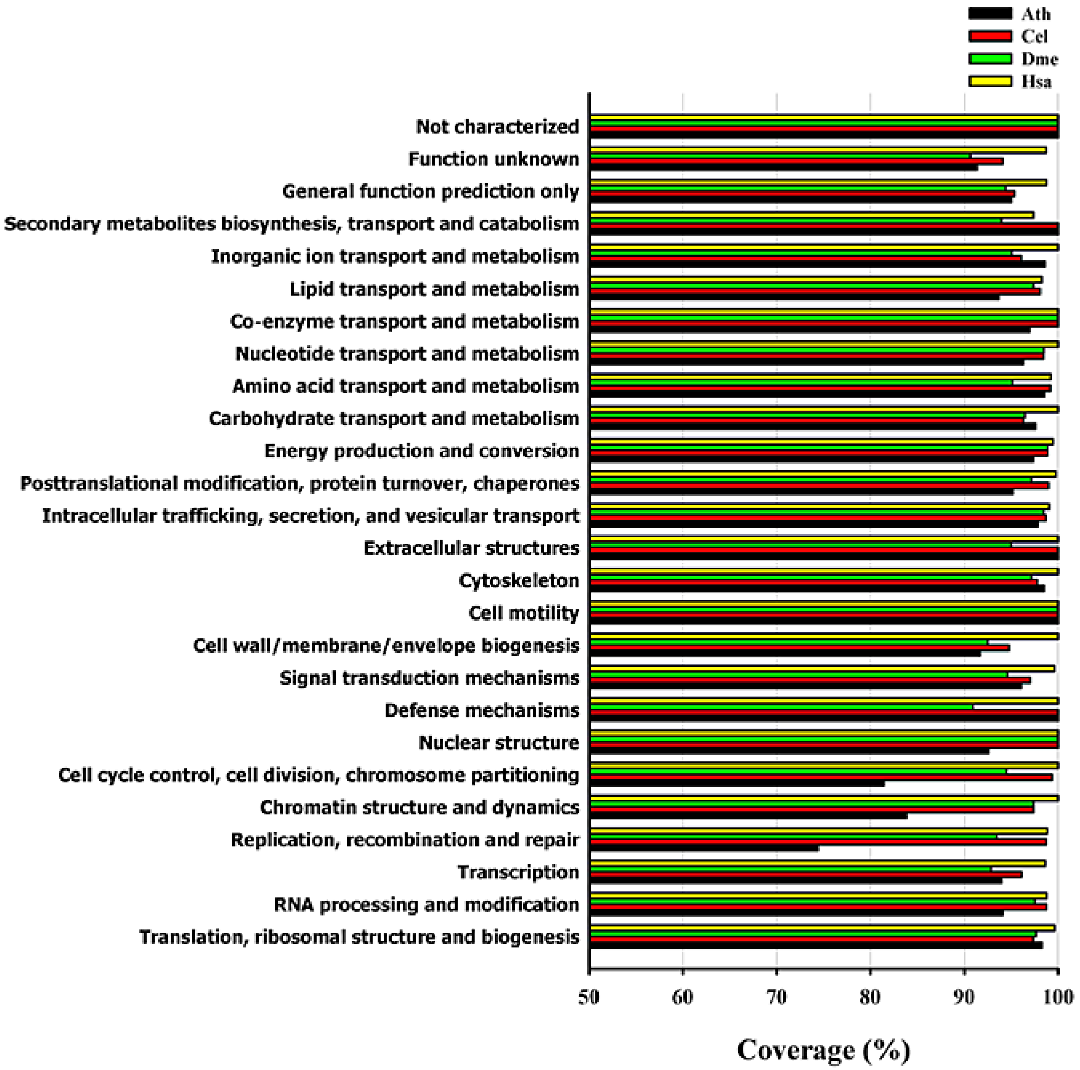

Figure 4. Coverage of KOG database (KOG, TWOG, LSE) by ESTs of the four eukaryotes (178,538, 215,200, 261,404, and 1,941,556 for Ath, Cel, Dme and Hsa, respectively). Ath = *Arabidopsis thaliana*; Cel = *Caenorhabditis elegans*; Dme = *Drosophila melanogaster*; Hsa = *Homo sapiens.*

potential, probably because the database contained shorter sequences (see supplementary Table S1 at http://biodados.icb.ufmg.br). In the second experiment (Figure 5B), the organisms' proteins were removed from the database when the cognate organisms' ESTs were annotated (e.g., ESTs from Dme would be annotated only by the other organisms' proteins, except Dme proteins). As expected, less effective coverage (around 10-20% loss compared to Figure 5A) was obtained. Figure 5B shows that Ath proteins were less efficiently covered by the incremental EST collections than Ath clusters. This behavior was expected, as it is the only plant in the database. Dme and Cel seemed not to be as affected in this second experiment, probably because their proteomes are relatively more related.

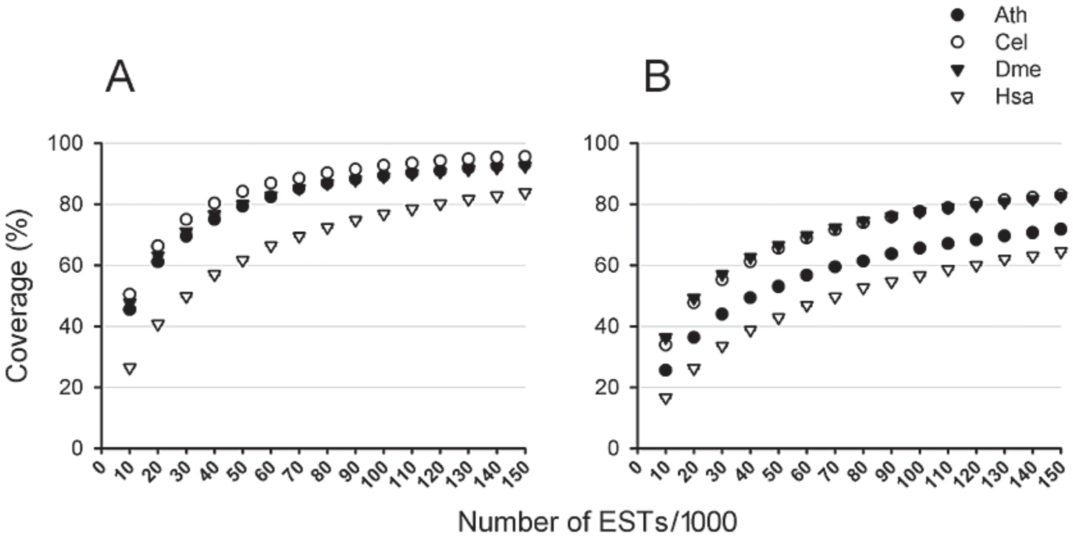

Figure 5. KOG coverage calculated by using 10 to 150 K ESTs (N = 10) from the four eukaryotes. **A.** Annotation with all proteins. **B.** Annotation with the cognate proteins depleted from the database. The standard error of the mean was under 1% in all events. Ath = *Arabidopsis thaliana*; Cel = *Caenorhabditis elegans*; Dme = *Drosophila melanogaster*; Hsa = *Homo sapiens*.

Figure 6 shows the same data from the last experiment, but with coverage distributed by KOG functional category. As seen in Figure 6A and B, Cel, Dme and Ath had similar patterns of KOG functional category coverage when using 10, 50, 100, and 150 K ESTs. By using unpaired *t*-tests with the 50 and 100 K coverage data, all organisms gave P values lower than 0.01. When the same test was run with the 100 and 150 K categories, Cel, Dme and Ath gave P values above 0.05. It is possible that after the exponential phase of the coverage curve (Figure 5A and B) producing more ESTs would be less effective in covering the KOG clusters, as these two coverage sets (100 and 150 K) were not statistically different. Hsa did not show this characteristic, as expected by the latter experiment. ‘Cell Motility’ and ‘Not categorized’ categories gave larger error bars since they were composed of small numbers of clusters (see supplementary Table S3).

The coverage of proteins from an organism, considering only KOG proteins (not TWOG or LSE), was analyzed (Figure 7). Figure 7A shows the protein coverage by the cognate organisms’ ESTs. Less than 50% of all proteins were covered by using up to 150 K ESTs. This happens because only a few paralog representatives of the clusters were being preferentially sampled (data not shown). Moreover, the coverage seemed to saturate when more ESTs were used, with the same exponential phase, followed by a linear plateau (Figure 5A). As little as 20% of the Ath KOG proteins were being covered, probably because of the large number of duplicated genes in this plant; few of them were sampled by the ESTs (data not shown). Hsa again covered fewer proteins than for the other organisms, possibly due to the smaller size of the ESTs added to putatively larger 3’ and 5’ UTRs. Figure 7B shows the quantity of KOG proteins covered when the species from which the ESTs originated was excluded. Comparing Figures 5B and 7B, it appears that no more than 6% of the non-cognate organisms’ proteins

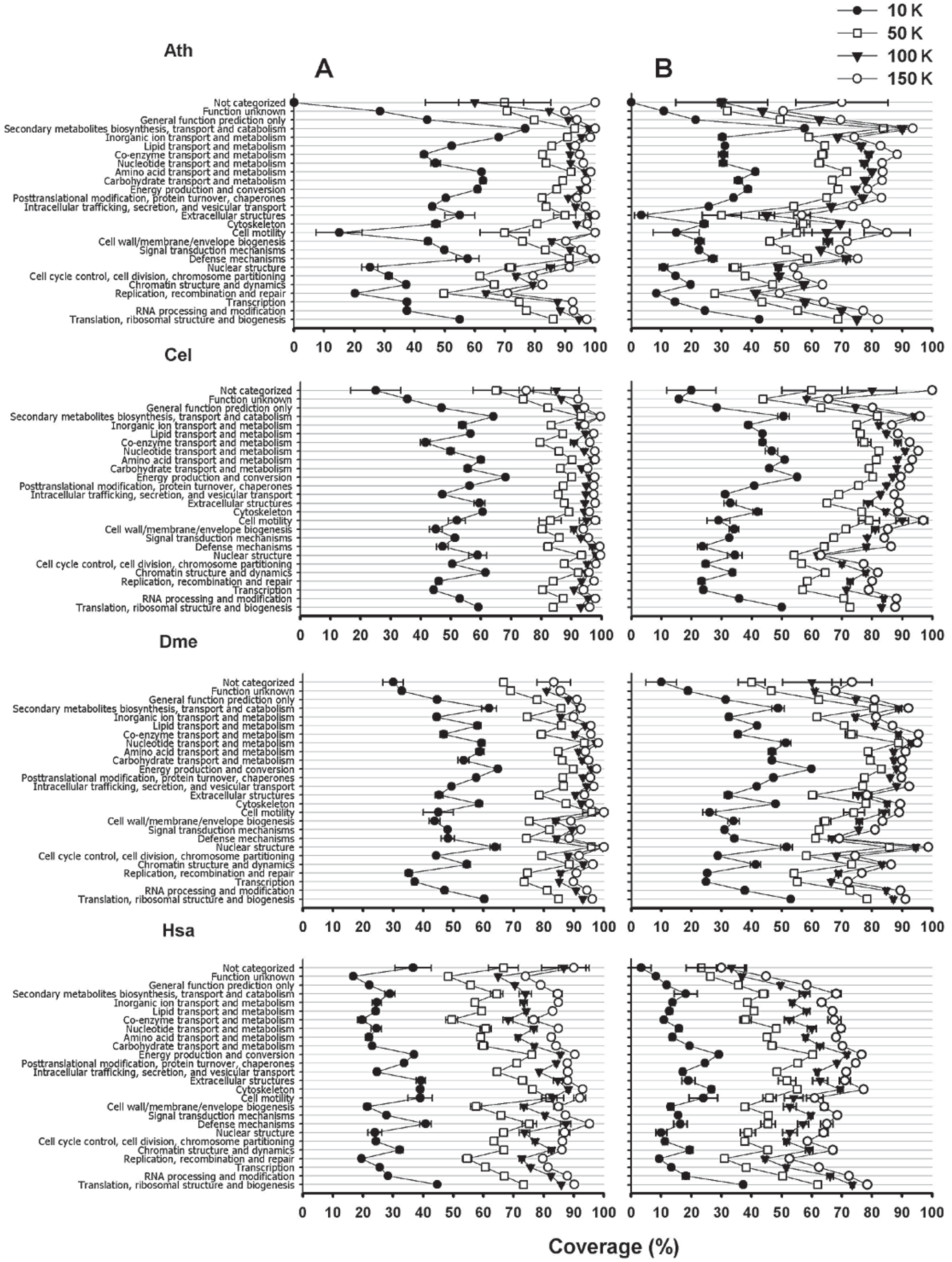

Figure 6. Coverage of KOG functional categories by using 10, 50, 100, and 150 K ESTs (N = 10) from the four eukaryotes. **A.** Annotation with all proteins. **B.** Annotation with the cognate proteins depleted from the database. Ath = *Arabidopsis thaliana*; Cel = *Caenorhabditis elegans*; Dme = *Drosophila melanogaster*; Hsa = *Homo sapiens.*

cover up to 60-80% of the KOG clusters. This shows that the KOG database does not entirely depend on the presence of proteins of the same organism, from which the ESTs originate, to annotate its ESTs, if KOG clusters are used instead of individual protein entries.

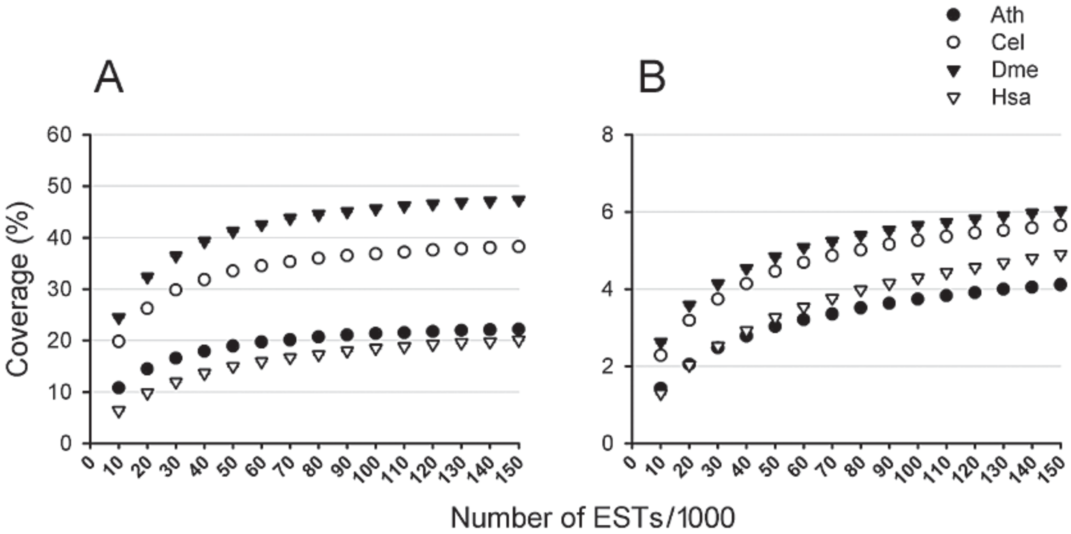

Figure 7. KOG protein coverage calculated by using 10-150 K ESTs (N = 10) from the four eukaryotes. **A.** Database including all proteins. **B.** Database without the cognate proteins. The standard error of the mean was under 1% in all cases. Ath = *Arabidopsis thaliana*; Cel = *Caenorhabditis elegans*; Dme = *Drosophila melanogaster*; Hsa = *Homo sapiens.*

CONCLUSIONS

The KOG protein coverage and cluster coverage results give a good indication on how to conduct a transcriptome project efficiently. Researchers can make predictions on how many ESTs should be produced in order to determine the genes that are most and least commonly expressed in other species. Results presented by each KOG entry can be accessed online in K-EST.

The study of sampling/expression of genes by ESTs with secondary databases generates answers to questions such as how many reads a transcriptome project should generate to cover a reasonable number of genes, or how frequently specific genes are expected to be sampled within a transcriptome project, given their sampling in model organism transcriptomes.

As the KOG database grows and incorporates more organisms, broader answers may be generated to these questions. In addition, new secondary databases are being developed, with more sequences and different organisms. The UniProt database (Bairoch et al., 2005) is such an example that we are currently investigating.

ACKNOWLEDGMENTS

Research supported by CAPES, FAPEMIG and CNPq/MCT.

REFERENCES

Adams MD, Kelley JM, Gocayne JD, Dubnick M, et al. (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* 252: 1651-1656.

Bairoch A, Apweiler R, Wu CH, Barker WC, et al. (2005). The Universal Protein Resource (UniProt). *Nucleic Acids Res*. 33: D154-D159.

Boguski MS, Lowe TM and Tolstoshev CM (1993). dbEST-database for “expressed sequence tags”. *Nat*. *Genet*. 4: 332-333.

Camon E, Barrell D, Brooksbank C, Magrane M, et al. (2003). The Gene Ontology Annotation (GOA) project: Application of GO in SWISS-PROT, TrEMBL and InterPro. *Comp*. *Funct*. *Genom*. 4: 71-74.

Ewing RM, Ben KA, Poirot O, Lopez F, et al. (1999). Large-scale statistical analyses of rice ESTs reveal correlated patterns of gene expression. *Genome Res*. 9: 950-959.

Faria-Campos AC, Cerqueira GC, Anacleto C, de Carvalho CM, et al. (2003). Mining microorganism EST databases in the quest for new proteins. *Genet*. *Mol*. *Res*. 2: 169-177.

Franco GR, Rabelo EM, Azevedo V, Pena HB, et al. (1997). Evaluation of cDNA libraries from different developmental stages of *Schistosoma mansoni* for production of expressed sequence tags (ESTs). *DNA Res*. 4: 231-240.

Hillier LD, Lennon G, Becker M, Bonaldo MF, et al. (1996). Generation and analysis of 280,000 human expressed sequence tags. *Genome Res*. 9: 807-828.

Kanehisa M and Goto S (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. 28: 27-30.

Lee NH, Weinstock KG, Kirkness EF, Earle-Hughes JA, et al. (1995). Comparative expressed-sequence tag analysis of differential gene expression profiles in PC-12 cells before and after nerve growth factor treatment. *Proc*. *Natl*. *Acad*. *Sci*. *USA* 92: 8303-8307.

Lindlof A (2003). Gene identification through large-scale EST sequence processing. *Appl*. *Bioinformatics* 2: 123-129.

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.

Mudado MA, Barbosa-Silva A, Torres JA, Paula-Pinto S, et al. K-EST: KOG Expression Sampling Tool. *Bioinformatics* (submitted).

Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, et al. (2001). The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res*. 29: 22-28.

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, et al. (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4: 41.