BIOINFORMATICS

Vol. 00 no. 00
Pages 1–2

K-EST : KOG Expression / Sampling Tool

Maurício A. Mudado, Adriano Barbosa-Silva, Saulo A.P. Pinto, João Torres, J. Miguel Ortega*

Departamento de Bioquímica e Imunologia, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, 31270-010, Belo Horizonte, MG, Brazil

ABSTRACT

Summary: K-EST is a web-based application that shows the annotation of large sets of ESTs from four model organisms (*A. thaliana*, *C. elegans*, *D. melanogaster* and *H. sapiens*) with the KOG database. K-EST may be used as a tool to predict the EST sampling or, roughly, gene expression in novel transcriptome projects by comparison to the four model organism expression / sampling database calculated with the KOG dataset. K-EST uses statistical methods to analyze differential expression between organisms and internal cDNA libraries. Other important feature of K-EST is to show the conservation between genes of the four organisms used.

Availability: http://www.biodados.icb.ufmg.br/K-EST/

Contact: miguel@icb.ufmg.br

Transcriptome projects aim to sample cell transcripts and complement genome projects as a support for gene mapping. The digital representation of sequences generated in these projects is known as EST or expressed sequence tags. Besides bearing up to 4% of sequencing errors, ESTs are intended to allow the identification of the codified protein throughout similarity searches (Adams *et al*., 1991). However, when a large number of ESTs originated from several independent cDNA libraries are available, they can be used to estimate gene expression (Lee *et al*., 1995, Franco *et al*., 1997). For the skeptic observer, EST occurrence provides, at least, an estimative of a given chance of gene sampling in an EST-based gene discovery program. Secondary databases, a repository of curate biological sequences, are a good source of information for annotation of novel sequences such as ESTs. The KOG (Eukaryote Cluster of Orthologs Group) database (Tatusov *et al*., 2003) is an interesting example for this purpose since its protein entries are classified into functional categories and groups. Moreover, KOG congregate into clusters, the proteins that exert analogous function in several model organisms with complete genome sequenced. These clusters represent one gene or protein that are therefore conserved during evolution time (e.g. enolase, represented by the ID KOG0047). In this work we present a tool that enables one to see the sampling of ESTs in four model organisms (Ath - *Arabidopsis thaliana*; Cel - *Caenorhabditis elegans*; Dme - *Drosophila melanogaster* and Hsa - *Homo sapiens*) that were annotated with the KOG database proteins using BLAST similarity searches (Altschul *et al*., 1990). It not only shows difference in expression / sampling between these organisms, but also allows one to evaluate whether some genes would appear or not in a novel transcriptome project by simply comparing EST sampling throughout organisms.

*to whom correspondence should be addressed

K-EST

KOG Expression / Sampling Tool is a web-based application that helps to predict EST sampling in novel transcriptome projects. It was developed using PHP (hypertext preprocessor) and relational database (MySQL), populated with BLAST results (best hits only, 10^{-10} E-value cutoff) from large sets of ESTs (Ath - 178,538; Cel - 215,200; Dme - 261,404 and Hsa - 1,941,556; for more information on sequences used see K-EST help page) queried against proteins from the KOG database. BLAST results were processed to depict the EST sampling by each KOG entry and or its associated functional category. The K-EST homepage was assembled to allow the user to select the combinations of 1, 2, 3 or 4 organisms and its respective expression / sampling within combinations of KOG functional categories, groups or KOG entries. All expression / sampling data were normalized by 100K ESTs or by generally more expressed transcripts like GAPDH and Actin, allowing to predict whether or not a gene would be sampled in novel transcriptome projects of different sizes. A tool was implemented to performs real-time annotations with entered query sequences against the KOG database and reports the sampling of the homologous genes. K-EST is also able to answer how ESTs assigned to a given KOG entry, for one organism, are being annotated by the other organism proteins; it is possible to verify how conserved is the chance of sampling a gene in one organism by using the complementary EST sampling information from the other organisms in K-EST conservation page.

EST SAMPLING

A simple method used to analyze the organism's EST sampling was to calculate the fold of expression, by dividing the highest sampling by the lowest. Another more reliable strategy used investigate differential expression between multiple cDNA libraries (Stekel *et al*., 2000). Its output is a single real value, called R, in which values above a threshold suggests differential expression between libraries of a given gene. This method was adapted in order to calculate difference in expression between the collections of ESTs, instead of cDNA libraries, by considering the set of ESTs of an organism as a single library. Although, the original method was also used in order to show if KOG clusters were also differentially expressed between cDNA libraries from specific organisms; differences are shown with different result cell colors (green, gold and red representing low, even and great differential expression) in the webpage tables.

CONSERVATION

BLAST searches were performed with the proteins and ESTs originated from cognate organisms (e.g. ESTs and proteins from Dme

Table 1. Examples of differential expression/sampling and conservation that exists simultaneously in different combination of organisms: Ath, Cel, Dme and Hsa, respectively represented by A, C, D and H.

Combination of organisms	% KOGs with R value ≤ 2	% KOGs with R value > 12	Cat. R ≤ 2* [fold of cat.]	Cat. R > 12* [fold of cat.]	% KOGs with iR value ≤ 12	% KOGs with Conserv. $\geq 80\%$
A C D H	324 / 2 542 (12.7%)	1 262 / 2 542 (49.65%)	S (21.2%) [2.48]	Q (86.36%) [13.4]	57 / 2 523 (2.25%)	214 / 2 279 (9.39%)
A C D -	915 / 2 587 (35.3%)	496 / 2 587 (19.17%)	S (56.1%) [2.48]	W (66.67%) [15.22]	528 / 2 572 (20.50%)	277 / 3 652 (7.58%)
A C - H	687 / 2 716 (25.3%)	896 / 2 716 (32.99%)	K (37.2%) [2.32]	Q (77.78%) [13.4]	194 / 2 702 (7.17%)	231 / 2 483 (9.30%)
A - D H	723 / 2 814 (25.7%)	1 028 / 2 814 (36.53%)	L (44.5%) [2.08]	Q (83.33%) [13.4]	77 / 2 794 (2.75%)	260 / 2 558 (10.16%)
- C D H	897 / 3 970 (22.6%)	1 483 / 3 970 (37.36%)	S (31.7%) [2.48]	E (60.36%) [5.39]	157 / 3 951 (3.97%)	417 / 3 652 (11.42%)
A C - -	1 698 / 2 793 (60.0%)	246 / 2 793 (8.81%)	S (75.3%) [1.77]	W (50.00%) [15.22]	2 053 / 2 785 (73.7%)	312 / 2 567 (12.15%)
A - - H	1 830 / 3 146 (58.1%)	480 / 3 146 (15.26%)	N (100.0%) [15.53]	Q (64.52%) [13.4]	273 / 3 121 (8.74%)	285 / 2 919 (9.76%)
A - D -	1 731 / 2 876 (60.2%)	311 / 2 876 (10.81%)	S (77.4%) [1.98]	Q (44.00%) [2.93]	643 / 2 860 (22.48%)	381 / 2 618 (14.55%)
- C - H	2 045 / 4 297 (47.6%)	790 / 4 297 (18.38%)	S (55.6%) [2.22]	E (40.50%) [5.39]	429 / 4 283 (10.01%)	610 / 4 114 (14.83%)
- - D H	2 178 / 4 706 (46.3%)	1 027 / 4 706 (21.82%)	L (62.4%) [2.08]	C (43.82%) [3.1]	246 / 4 680 (5.25%)	606 / 4 397 (13.78%)
- C D -	2 599 / 4 106 (63.3%)	313 / 4 106 (7.62%)	N (100.0%) [1.06]	Q (20.69%) [1.07]	1 105 / 4 091 (27.01%)	865 / 3 135 (22.82%)

* Category with higher percentage of KOGs with indicated R value. The category X (not categorized) was excluded.

only), to reveal the actual EST sampling of each organism. Thereafter, the cognate organisms proteins were removed and only the proteins from the other organisms were used (e.g. ESTs from Dme and proteins from Ath, Cel and Hsa) to show whether the ratio of EST sampling was maintained or the sampling was diminished (indicating that KOG proteins would be less conserved between organisms). The level of conservation is depicted by colors (green, gold and red, representing above 80%, between 20-80% or fewer than 20% of conservation). It is supposed that a direct comparison between KOG proteins might conduct to equivalent results, although the results shown in K-EST are operational.

Table 2. Conservation of KOG and KOG categories

KOGs	Ath	Cel	Dme	Hsa
Well conserved	11.14%	24.26%	42.64%	12.60%
Poorly conserved	55.77%	36.51%	26.90%	42.06%
Most conserved[a]	J (32.4%)	C (60.4%)	J (73.5%)	J (39.2%)
Less conserved[b]	W (90.0%)	Y (50%)	W (33.3%)	V (66.2%)

a,b. Percentage of KOGs in the category that are well and poorly conserved, respectively. Category X was excluded.

EXAMPLES OF DATA MINED FROM K-EST

Table 1 shows interesting examples of data mined from K-EST, like the percentage of KOGs shared by the combination of the four organisms with an even expression (R value lower than or equal to 2) or with a differential expression (R value higher than 12). Also the same approach was used to KOG categories, internal R values (R values calculated from cDNA libraries from specific organisms) and conservation. Interesting information can be extracted from this analysis, like the two organisms that share more KOGs evenly expressed and less KOGs differentially expressed (Cel and Dme - 63% and 7.62% respectively). This can be interpreted by the shorter evolutionary distance between these two organisms, comparing to the others. Individually, Dme shows the highest frequency of conserved KOGs (KOGs with conservation above or equal to 80%, Table 2), 42.64%; surprisingly Cel does not follow this behavior (24.26%). Also, Ath presents 55.77% of poorly-conserved KOGs (KOGs below 20% conservation) as expected, since it is the only plant. Additional interpretations may be extracted by K-EST users.

ACKNOWLEDGEMENT

This work was supported by CAPES, FAPEMIG and CNPq.

REFERENCES

Adams,M.D. et al., (1991) Complementary DNA Sequencing: Expressed Sequence Tags and Human Genome Project. *Science*, **252**, 1651-1656.

Altschul,S.F. et al., (1990) Basic local alignment search tool. *J Mol Biol*, **215**, 403-10.

Franco,G.R. et al., (1997) Evaluation of cdna libraries from different developmental stages of schistosoma mansoni for production of expressed sequence tags (ests). *DNA Res*, **4**, 231-40.

Lee,N.H. 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 U S A* 8303-7, 1995.

Tatusov,R.L. et al., (2003) The cog database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.

Stekel,D.J., Git,Y., Falciani,F., (2000) The comparison of gene expression from multiple cDNA libraries. *Genome Res*, **10**, 2055-61.