



EPPS: mining the COG database by an extended phylogenetic patterns search

Klaus Reichard¹ and Michael Kaufmann^{2,*}

¹Institute of Mathematics, AG Statistics and ²Institute of Neurobiochemistry, AG Proteinchemistry, University of Witten/Herdecke, Stockumer Str. 10, 58448 Witten, Germany

Received on July 7, 2002; revised on November 8, 2002; accepted on December 14, 2002

ABSTRACT

Summary: EPPS runs under Microsoft Windows. It is an extended version of the phylogenetic patterns search (PPS). The output condition of PPS is the exact match of a user defined phylogenetic pattern with the pattern represented by the respective cluster of orthologous groups (COG). In contrast, the software described here is less restrictive. The user may define the accuracy of the search by the number of genomes that are allowed not to match the predefined phylogenetic pattern. Thus, EPPS has the advantage to detect COGs even if organisms defined to be included are not or organisms defined to be excluded are present in the output COGs.

Availability: EPPS is free and both windows executable and source code are available at <http://www.uni-wh.de/de/nawi/institut/protein/epps.html>

Contact: mika@uni-wh.de

Comparative genomics led to the definition of clusters of orthologous groups of proteins (COGs) by comparing protein sequences encoded in 43 complete genomes, representing 30 major phylogenetic lineages (Tatusov *et al.*, 1997, 2001). Each COG consists of individual proteins or groups of orthologs from at least three lineages and thus corresponds to a conserved domain. The COG database can be analyzed by a phylogenetic patterns search (PPS), a tool that provides a means for finding COGs that contain or exclude a selected organism (<http://www.ncbi.nlm.nih.gov/COG/phylox.html>). For that purpose, the user needs to define a pattern of the 30 lineages by assigning each lineage one of the following three parameters: dc=the COG may or may not contain this lineage, Yes=the COG must contain this lineage, No=the COG must not contain this lineage. The list that results after analysis is the subset of COGs that exactly fits the predefined pattern. Such a phylogenetic pattern search is a powerful tool to extract biological information from the COG database e.g. reverse gyrase was detected by PPS to be the only hyperthermophile-specific protein (Forterre,

2002). However, when performing PPS, COGs that do not exactly fit the input pattern may also be of biological relevance with respect to a certain question. For example, if COGs exclusively containing thermophilic proteins are searched, this could also be the case for those COGs not containing one or more thermophilic organisms. On the other hand, although a COG contains one or more organisms that were excluded by the user, such a COG may also be of interest. For that reason, we developed an extended phylogenetic patterns search (EPPS). Compared to PPS, this program considers not only the 30 lineages but all 43 organisms of the COG database and it has two additional input parameters. EPPS allows the researcher to specify the accuracy of the analysis with respect to (i) the organisms that must and (ii) the organisms that must not be included in the output COGs i.e. the maximum number of exceptions of an exact match for each case can be defined. To demonstrate the suitability of EPPS in comparative genomics, we analyzed COGs containing thermophilic (including hyperthermophilic) organisms. An exact phylogenetic pattern search (exceptions ≤ 0 for Yes and No, respectively) revealed only one COG (COG1618 Predicted ATPases or kinases). In contrast, allowing one exception for Yes, two additional COGs were output (COG1980 Uncharacterized ACR; COG1350 Predicted alternative tryptophan synthase beta-subunit). The importance of COG1350 with respect to thermophily was recently described (Hettwer and Sterner, 2002). A second analysis allowing one exception for No also revealed another COG (COG3635 Predicted phosphoglycerate mutase, AP superfamily). Further analysis showed, that in addition to all thermophilic organisms this COG contains a protein from *Deinococcus radiodurans*. Since *D.radiodurans* is a polyextremophile, showing remarkable resistance to a range of severe damage caused by ionizing radiation, desiccation, UV radiation, oxidizing agents, or electrophilic mutagens (Minton, 1994) it is very likely that its COG3635 protein is also thermophilic.

The software described here (EPPS.exe) runs as a stand alone application under Microsoft Windows. Two additional files, COGs.txt and org.txt, are required in

*To whom correspondence should be addressed.

the same folder for input. The application is capable to download those files via ftp (<ftp://ftp.ncbi.nih.gov/pub/COG/>) on user request or during the first execution. Provided those files remain in the present format, the package is capable to dynamically update its graphical user interface even if more than the present 43 microbial genomes will be included in the future.

REFERENCES

- Forterre,P. (2002) A hot story from comparative genomics: reverse gyrase is the only hyperthermophile-specific protein. *Trends Genet.*, **18**, 236–237.
- Hettwer,S. and Sterner,R. (2002) A novel tryptophan synthase beta-subunit from the hyperthermophile *Thermotoga maritima*. Quaternary structure, steady-state kinetics, and putative physiological role. *J. Biol. Chem.*, **277**, 8194–1201.
- Minton,K.W. (1994) DNA repair in the extremely radioresistant bacterium *Deinococcus radiodurans*. *Mol. Microbiol.*, **13**, 9–15.
- Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.
- Tatusov,R.L., Natale,D.A., Garkavtsev,I.V., Tatusova,T.A., Shankavaram,U.T., Rao,B.S., Kiryutin,B., Galperin,M.Y., Fedorova,N.D. and Koonin,E.V. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.*, **29**, 22–28.