3720–3722 *Nucleic Acids Research, 2003, Vol. 31, No. 13*
DOI: 10.1093/nar/gkg603

Phydbac (phylogenomic display of bacterial genes): an interactive resource for the annotation of bacterial genomes

François Enault*, Karsten Suhre, Olivier Poirot, Chantal Abergel and Jean-Michel Claverie

Structural and Genomic Information (CNRS UPR2589), Institut de Biologie Structurale et Microbiologie,
31 chemin Joseph Aiguier, 13402 Marseille, cedex 20, France

Received February 15, 2003; Revised and Accepted March 31, 2003

ABSTRACT

Phydbac is a web interactive resource based on phylogenomic profiling, designed to help microbiologists to annotate bacterial proteins. Phylogenomic annotation is based on the assumption that functionally linked protein-coding genes must evolve in a coordinated manner. The detection of subsets of co-evolving genes within a given genome involves the computation of protein sequence conservation profiles across a spectrum of microbial species, followed by the identification of significant pairwise correlations between them. Many ongoing studies are devoted to the problem of computing the most biologically significant phylogenomic profiles and how best identifying clusters of 'functionally interacting' genes. Here we introduce a web tool, Phydbac, allowing the dynamic construction of phylogenomic profiles of protein sequences of interest and their interactive display. In addition, Phydbac can identify *Escherichia coli* proteins exhibiting the evolution pattern most similar to arbitrary query protein sequences, hence providing functional hints for open reading frames (ORFs) of hypothetical or unknown function. The phylogenomic profiles of all *E.coli* K-12 protein-coding genes are pre-computed, allowing queries about *E.coli* genes to be answered instantaneously. The profiles and phylogenomic neighborhoods are computed using an original method shown to perform better than previous ones. An extension of Phydbac, including precomputed profiles for all available bacterial genomes (including major pathogens) will soon be available. Phydbac can be accessed at: http://igs-server.cnrs-mrs.fr/phydbac/.

INTRODUCTION

Determining protein functions from genomic sequences is one of the main challenges of bioinformatics. To this purpose, alignment methods based on sequence similarity, such as PSI-BLAST (1) or Pfam (2), are the most heavily used and are still being refined. Yet, they are only capable of providing reliable functional predictions for ~50% of the open reading frames (ORFs) of most newly sequenced microorganisms (3), corresponding to the proportion of already functionally annotated protein coding genes. Besides the experimental determination of gene functions, escaping this vicious circle requires the development of bioinformatic approaches going beyond the recognition of sequence similarity and functional signatures. Phylogenomic profiling is one of these new methods. It is based on the assumption that proteins involved in a common metabolic pathway or constituting a multi-molecular complex are likely to evolve in a correlated manner. We use the term co-evolution throughout this paper to designate such a behavior. This paradigm, originally named phylogenetic profiling, was first put to use by Pellegrini *et al.* (4) who demonstrated that some information on the function of a protein could be retrieved by analyzing the functions of its phylogenomic neighbors, this neighborhood being defined as the subset of the best co-evolving genes in the same genome (e.g. *Escherichia coli*). This approach has been subsequently used in many studies (5–7) and the definition of a meaningful neighborhood refined in various ways (8–10). Other phylogenomic methods have been proposed, such as the analysis of gene co-localization (11,12), as well as the systematic search for gene fusion events (13,14). At the moment, Phydbac only uses the initial concept of co-evolution, but the co-localization information will be added in the future.

In the current version, Phydbac emulates two main different modes of operation. In the first mode, the software allows researchers to build, display and compare the evolution profile(s) of their protein(s) of interest, for instance to see if they exhibit any evidence of co-evolution. The phylogenomic profile is computed on line, using an ORF database derived from 71 bacterial and archaeal (non-redundant) species. By analyzing each sequence conservation profile individually and/or

*To whom correspondence should be addressed. Tel: +33 491164548; Fax: +33 491164549; Email: enault@igs.cnrs-mrs.fr

Published by Oxford University Press 2003

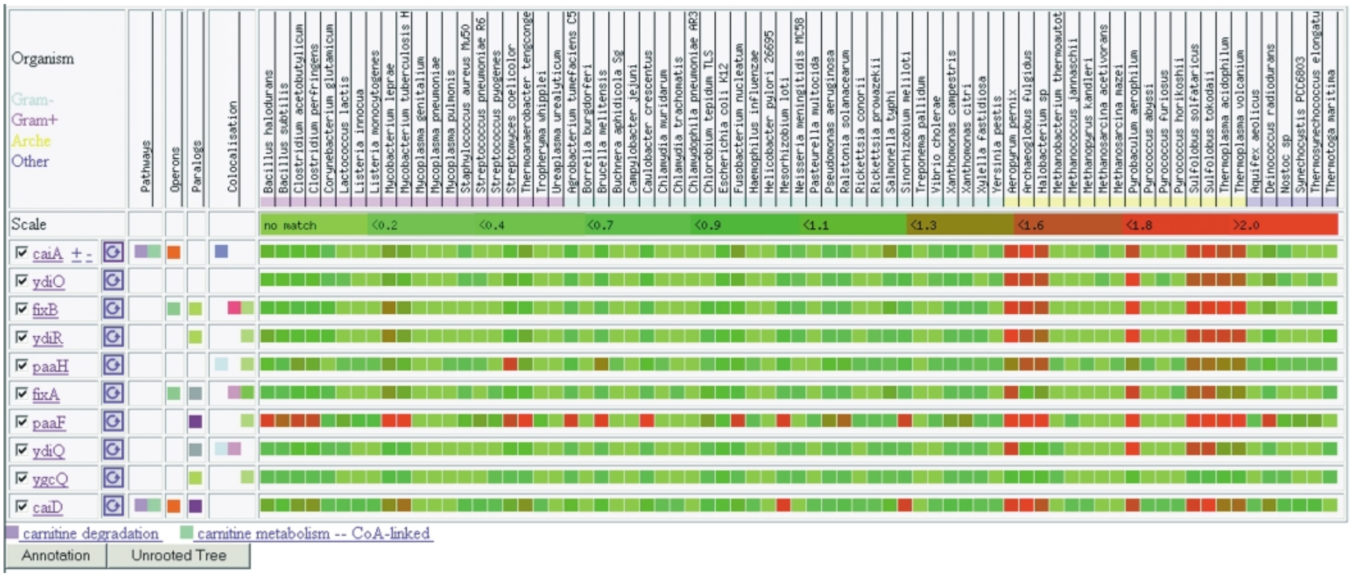

Figure 1. Typical output of a standard Phydbac session. Profiles display of the phylogenomic neighbors of the gene caiA.

comparing them between genes, one may infer hypotheses on their function. To take a simplistic example, an ORF only found to be conserved in bacteria with flagella might be suspected to have a role in motility. Obvious signs of co-evolution between query proteins can be detected by simply visualizing their conservation profiles. Phydbac profiles are most useful to corroborate (or invalidate) biologist intuitions when applied to proteins already suspected to be functionaly linked from previous—albeit not entirely convincing—evidence.

Phydbac's second mode of operation is restricted to all previously defined *E.coli* ORFs. Phylogenomic neighborhoods can be instantaneously displayed for all of them, using our improved definition of pairwise gene distance (10) computed from the correlations of conservation profiles. This mode is used to generate functional hypotheses about the numerous *E.coli* genes still remaining anonymous. To complement this approach, we implemented a BLAST search tool allowing an arbitrary protein sequence query to be associated to its homolog in *E.coli*. This can provide a starting point to build functional hypotheses, based on sequence and/or phylogenomic profile similarities with *E.coli* homologs.

METHODS

In order to build phylogenomic profiles, we compare the query sequences to all ORFs from 71 non-redundant (only one strain per species is used) bacterial and archaeal genomes using BLASTP (1). Each point in the phylogenomic profile reflects the similarity between the query protein and its best matching ORF within each of the 71 genomes (each one corresponding to a fixed column in the plot). More precisely, its value is the largest BLAST bit score of the alignment between the query protein and all ORFs of the given genome, divided by the self-alignment score of the query protein. The self-alignment being the best scoring one, the profile values (called normalized score) span a [0–1] range. This allows each point to be weighed proportionally to the length and quality of the alignment independently of the total protein length. For the analysis of *E.coli* K-12 genes, we selected 4263 of the 4279 known ORFs longer than 50 amino acids and applied the above protocol to each of them. A second normalization procedure was then used on the resulting 4263 profiles to compensate for the decreasing protein similarity (i.e. relative BLAST score) expected when comparing homologous genes from organisms at increasing evolutionary distance. Each profile column (i.e. each genome) was normalized by the average of the non-zero normalized BLAST scores (i.e. above the bit score threshold) obtained for this organism. In a separate study (10) we evaluated the performance of different phylogenomic pairwise distances between genes computed from their conservation profiles. The best performing method, using the Ecocyc database (15) as a reference, was used in Phydbac to define the phylogenomic neighborhood of each *E.coli* ORF.

WEB INTERFACE

Phydbac is an interactive web resource accessible at http://igs-server.cnrs-mrs.fr/phydbac/. The different options available through Phydbac's main page reflects its different modes of operation. The first one inputs a single file of fasta-formatted protein sequences to dynamically create their conservation profiles. Under the current hardware implementation, building a single protein profile requires 5 s. This involves a BLASTP comparison against a one million ORF database and the generation of the profile graphics. Thus, about 25 query sequences is the limit for an interactive session. Fortunately one is rarely interested by comparing more than a handful of genes at a time. An option allows the phylogenomic relationships to be displayed as a tree. This unrooted tree is built by applying the neighbor-joining method to the phylogenomic pairwise distance matrix. The second operation mode provides a direct access to *E.coli* genes sequence conservation profiles and phylogenomic neighbors. *E.coli* genes can be retrieved by their names, the presence of a keyword in their annotation or by similarity with a user-provided query sequence. Upon the selection of one or several genes, their profiles are displayed (Fig. 1). Clicking on the icon near the name of the genes identifies its 10 closest phylogenomic neighbors and gives access to their conservation profiles. As the number of neighbors is arbitrary, the possibility of getting more or less than 10 neighbors is offered (using a plus or minus button near the query's name).

In addition to the profiles, accessory information is also displayed for any gene list (manually selected genes or a query and its neighbors). When two or more genes of a list belong to the same pathway, to the same operon, have a significant sequence similarity or are found to be co-localized, it is indicated in the four corresponding columns, right before their profiles. For instance, Figure 1 shows that caiA and caiD are involved in common pathways (carnitine degradation and carnitine metabolism—CoA-linked) and that they belong to the same operon. The 'Paralogs' column shows that fixB, ydiR and ygcQ share some significant sequence similarity. Finally, the 'Colocalization' column indicates that paaH and ydiC are co-localized with caiA (i.e. their respective homologous sequences are separated by <2000 nucleotides in more than three genomes). As co-localization is not a transitive property (paaH and ydiC are not co-localized), each gene found co-localized with other genes in the list generates its own column of colored markers, the darkest hue being associated with each reference gene. For selected genes (checkboxes are on the left of the gene names), annotations and an unrooted tree, built as described before, can be displayed in pop-up windows. Finally, expression intensity values (e.g. from DNA-chip experiments) can be loaded in parallel and displayed next to the profiles of the selected genes (or any of its neighbors). The visual comparison of phylogenomic versus expression intensity profiles may help in the generation of hypotheses on putative functions and metabolic processes.

FUTURE PROSPECTS

Future versions of Phydbac will extend the mode of operation currently limited to *E.coli* ORFs to all fully sequenced micro-organisms. Priority will be given to major human pathogens. This will require the integration/incorporation of the functional annotations contained in databases on different bacterial genomes and the storage of the profiles of all bacterial ORFs.

ACKNOWLEDGEMENT

Thanks to B. Chagot for suggesting the catchy 'phydbac' name and for helpful discussions.

REFERENCES

1. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
2. Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L., Eddy,S.R., Griffiths-Jones,S., Howe,K.L., Marshall,M. and Sonnhammer,E.L. (2002). The Pfam protein families database. *Nucleic Acids Res.*, **30**, 276–280.
3. Claverie,J.M., Abergel,C., Audic,S. and Ogata,H. (2001) Recent advances in computational genomics. *Pharmacogenomics*, **23**, 61–72.
4. Pellegrini,M., Marcotte,E.M., Thompson,M.J., Eisenberg,D. and Yeates,T.O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA*, **96**, 4285–4288.
5. Marcotte,E.M., Pellegrini,M., Thompson,M.J., Yeates,T.O., Eisenberg,D. (1999) A combined algorithm for genome-wide prediction of protein function. *Nature*, **402**, 83–86.
6. Marcotte,E.M., Xenarios,I., van Der Bliek,A.M. and Eisenberg,D. (2000) Localizing proteins in the cell from their phylogenetic profiles. *Proc. Natl Acad. Sci. USA*, **97**, 12115–12120.
7. Pavlidis,P., Weston,J., Cai,J. and Noble,W.S. (2002) Learning gene functional classifications from multiple data types. *J. Comput. Biol.*, **9**, 401–411.
8. Zheng,Y., Roberts,R.J. and Kasif,S. (2002) Genomic functional annotation using co-evolution profiles of gene clusters. *Genome Biol.*, **3**, research0060.
9. Bilu,Y. and Linia,M. (2002) Functional consequences in metabolic pathways from phylogenetic profiles. *WABI 2002, Second International Workshop: Algorithms in Bioinformatics*, pp. 263–276.
10. Enault,F., Suhre,K., Poirot,O., Abergel,C. and Claverie,J.M. (2003) Annotation of bacterial genomes using improved phylogenomic profiles. *ISMB Supplement to Bioinformatics*, in press.
11. Overbeek,R., Fonstein,M., D'Souza,M., Pusch,G.D. and Maltsev,N. (1999) The use of gene clusters to infer functional coupling. *Proc. Natl Acad. Sci. USA*, **96**, 2896–2901.
12. Dandekar,T., Snel,B., Huynen,M. and Bork,P. (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.*, **23**, 324–328.
13. Marcotte,E.M., Pellegrini,M., Ng,H.L., Rice,D.W., Yeates,T.O. and Eisenberg,D. (1999) Detecting protein function and protein-protein interactions from genome sequences. *Science*, **285**, 751–753.
14. Enright,A.J., Iliopoulos,I., Kyrpides,N.C. and Ouzounis,C.A. (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**, 86–90.
15. Karp,P.D., Riley,M., Saier,M., Paulsen,I.T., Collado-Vides,J., Paley,S.M., Pellegrini-Toole,A., Bonavides,C. and Gama-Castro,S. (2002) The EcoCyc Database. *Nucleic Acids Res.*, **30**, 56–58.