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/.
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 functionally 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.
ACKNOWLEDGEMENT

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

REFERENCES


