PublisherInfo		
PublisherName	:	BioMed Central
PublisherLocation	:	London
PublisherImprintName	:	BioMed Central

Phylogenetic classification of proteins encoded in complete genomes

ArticleInfo		
ArticleID	:	3613
ArticleDOI	:	10.1186/gb-2000-1-1-reports239
ArticleCitationID	:	reports239
ArticleSequenceNumber	:	104
ArticleCategory	:	Web report
ArticleFirstPage	:	1
ArticleLastPage	:	4
ArticleHistory	:	RegistrationDate : 2000–2–29 Received : 2000–2–29 OnlineDate : 2000–4–27
ArticleCopyright	:	BioMed Central Ltd2000

ArticleGrants	:	
ArticleContext	:	130591111

Todd Richmond

Abstract

For those interested in genome-wide or more restricted comparisons of proteins across species, the Clusters of Orthologous Groups (COGs) website provides the kind of information many of us want.

Content

For those interested in genome-wide or more restricted comparisons of proteins across species, the Clusters of Orthologous Groups (COGs) website provides the kind of information many of us want but for which we lack the necessary computational tools and/or power. For 21 completely sequenced organisms, including *Escherichia coli*, *Saccharomyces cerevisiae* and *Haemophilus influenzae*, the site gives all the clusters of paralogous proteins shared by three or more lineages. Currently, 2,112 COGs are listed. This database can be searched in many different ways - for proteins shared among particular organisms, for proteins found in some organisms but not others, for proteins found in particular metabolic pathways or by functional classification. The site provides more information than a simple BLAST search for sequence similarity can, especially for bacterial proteins. In bacteria, genes have often been sequenced many times, and it is difficult to determine from a simple search whether a particular bacterium has one gene, sequenced five times, or five genes, each sequenced once. COG removes some of that ambiguity by using only completely sequenced and annotated genomes.

Navigation

The entry page contains a list of the 21 organisms and a series of links to the COGs, distributions, phylogenetic patterns, functional categories of genes and metabolic pathway information. Imagine you have cloned a gene and discovered a similarity to the *E. coli* gene encoding the protein FadB, an enzyme involved in fatty-acid beta-oxidation. To find out how common this gene is and which organisms have this particular protein, type in FadB on the COG page, and it returns the two domains associated with that protein: enoyl-CoA hydratase/carnithine racemase and 3-hydroxyacyl-CoA dehydrogenase. Clicking on the dehydrogenase COG takes you to a summary page. This page gives the number of proteins that match that description, the functional classification and metabolic pathway, links to the various sequences from each genome (showing which genomes have paralogs and which do not), a small phylogenetic tree, and a link to save all of the proteins to disk in FASTA format. All the

information is interconnected, so you can click on sequence to see the alignments and/or BLAST reports, or click on a pathway to see all the other COGs in that pathway. It is possible to search for subsets of genes that are shared by all organisms, or by a subset of organisms that you designate. For example, a search for all predicted membrane proteins of uncharacterized function will give 54 COGs; a query asking which ones are found in both yeast and *E. coli* gives just one, COG0706 Inner membrane proteins, SpoIIIJ family. There is a separate page of co-occurences that indicates how many genes are globally shared by two organisms and how many are unique to each organism.

Reporter's comments

Timeliness

The site was last updated on 24 January 2000.

Best feature

There is a comprehensive help page which is essential for extracting the most information out of the site.

Worst feature

The information tends to be arranged in huge tables that take a long time to load. For example, to browse the list of COGs requires first loading a 431K table. When browsing through individual COGs, this table has to be reloaded each time. Ideally, the links should open in a new window so that the table never has to be reloaded. The color and abbreviations used throughout the site can be cryptic. Genomes are designated by one-letter codes, some of which are not obvious; E for *E. coli* and Y for *S. cerevisiae* (yeast) are fine, but R for *Mycobacterium tuberculosis* and I for *Chlamydia trachomatis*? Why not use a more intuitively obvious two-letter code? As for the colors, whoever designed the site must have had good color vision: but subtle shades are not the best method of conveying information for some people.

Wish list

Although there are advantages in restricting COGs to completed genomes, it would be nice if a few partially sequenced higher eukaryotes were included. Sometimes one wants to ask "Is this protein found in both eukaryotes and prokaryotes, and how divergent is it?"

Related websites

The COGs site is unique. Although there are others that collect information about protein domains and/or specific protein families, none of these allows browsing through the collected proteins from 21 complete genomes. It is debatable, however, how long this site will be able to maintain its present form. According to the Genomes online database, there are currently 107 prokaryotic and 31 eukaryotic genomes being sequenced, and in a few years the site will become too cumbersome in its present format.

Table of links

Clusters of Orthologous Groups

Genomes online database

References

1. Clusters of Orthologous Groups.