Scope of the Protein Clusters database

The Protein Clusters database [1] is a collection of Reference Sequence (RefSeq) proteins grouped and annotated based on sequence similarity and protein function. The sequences are from the annotated complete genomes of prokaryotes, plasmids, viruses, organelles, protozoa and plants, as well as certain incomplete genomes of protozoa and plants. Proteins from different taxonomic groups or different types of genomic locations are contained within separate cluster groups.

The clusters are divided into curated and non-curated sets. Non-curated clusters are automatically generated and have not yet been manually curated. The manual curation involves confirming protein domain structure, joining related clusters, addition of publications, and functional annotation such as addition of protein names and Enzyme Commission numbers. Cluster annotation is used in RefSeq protein annotation and the Prokaryotic Genome Automatic Annotation Pipeline (PGAAP) [2] for the annotation of prokaryotic genomes submitted to NCBI.

Accessing the data

Data from this database can be searched and retrieved using the Protein Clusters homepage (A). Shown below are three of the records (B) retrieved by searching with “lacZ” (C). The “Advanced” page (D) allows more detailed examination of indexing fields (E) and terms (F), as well as access to search histories (G) for building more specific and powerful searches.

Links to a Protein Cluster are also available from each protein member of a cluster (G).

Curated clusters are also mirrored in the CDD database and searchable using a query protein sequence through the rpsBLAST interface, which is conveniently linked from the “Specialized BLAST” section of the BLAST homepage.
Record display

Shown below is protein cluster PCLA_419626 for beta-D-galactosidase in its overview format, which provides a general summary on the function of the protein (A) with multiple sections below to provide information on specific aspects of the cluster. Information provided in the Statistics section (B) shows the level of conservation through the number of proteins contained in the cluster and genera and organisms it covers. Records from other NCBI databases related to this cluster are linked in the "Related Information" section (C) in the right hand column. Collapsible sections below the summary and Statistics provides additional details of the cluster. The "Related clusters" lists clusters under the same node of the cluster hierarchy and their summary statistics. The "Genome groups" lists Genome entries covered by the cluster. The "Protein Tables" lists all the proteins included in the cluster. The "Filters" uses custom input, such as the organism name (D) to filter the list of proteins displayed. By default, identical proteins are hidden from the display.
Cluster access from protein sequence

Protein Clusters annotation on protein records can be viewed in the graphical display of a protein record. The example below (A) shows the beta-D-galactosidase protein (NP_414878) from Escherichia coli str. K-12 MG1655, which maps domain annotation to specific protein coordinates (B). The summary of a domain can be viewed by mouse-over (C) with the example popup representing the mirrored cluster PCLA_419626 in CDD. The “Identify Conserved Domains” link (D) provides similar information by submitting the protein sequence to rpsBLAST to identify conserved domains present in the sequence. Since the example protein is a member of the protein cluster PCLA_419626, a link to the protein cluster (E) along with the summary information for this cluster is shown in the right-hand column.

Clicking the “Identify Conserved Domains” link (F) opens up a rpsBLAST (CDD) search result page (below), which displays the different domains identified in this protein record. A mouse-over of an identified domain, lacZ in this case, shows the summary for that domain (G, the mirrored PRK09525). The alignment display can be toggled using the “+” and “-” signs (H), which expands to show the sequence alignments and selected BLAST statistics.
Access Protein Clusters using Concise BLAST

The Concise Protein BLAST interface (A) provides access to a consolidated set of sequence data from Protein Clusters. This database consists of all curated and non-curated protein clusters as well as non-clustered proteins, with each cluster sliced at the level of genera to represent “subclusters.” A single sequence is selected in random from each subcluster as its representative. This reduces the level of redundancy, leads to speedier searches, but still provides a broader taxonomic view than is typically found in protein BLAST results. This makes it possible to search the Protein Clusters data using a protein or nucleotide sequence as the query, with blastp and blastx programs (B), respectively.

The Concise BLAST result for a hypothetical protein from Arabidopsis is shown below. Here, the query and search summary are given at the top (C). All top hits belong to PCLA_2502852, a curated cluster for family 2 glycoside hydrolase (D). Toggling the “+” opens the cluster (E) showing the “collapsed” structure of the database: YP_00391530, with score (F), is the “single representation” at the genus-level for Arabidopsis, with other entries from the cluster (without score) displayed close to the representative. Note that a Concise BLAST result can also be displayed in traditional BLAST output using the “show results in standard format” link (G) given at the top of the page.

Reference, help documentation and FTP