



Protein Clusters Database

Grouping of selected proteins by sequence similarity and function
<http://www.ncbi.nlm.nih.gov/proteinclusters>

National Center for Biotechnology Information • National Library of Medicine • National Institutes of Health • Department of Health and Human Services

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.

Display Settings: Overview **Send to:**

PCLA_419626 beta-D-galactosidase ID: 419626

forms a homotetramer; hydrolyzes lactose disaccharide to galactose and glucose; converts lactose to allolactose which is the natural inducer of the lac operon

Statistics

Proteins: 153 (111 identical groups)
 Conserved in: Bacteria
 Total genera: 19
 Total organisms: 143
 Putative Paralogs: 2
 Locuses: lacZ, lacZ3
 EC Number: 3.2.1.23
 CDDs: pfam02929:Bgal_small_N(superfamily smart01038:Bgal_small_N(superfamily pfam00703:Glyco_hydro_2(superfamily pfam02836:Glyco_hydro_2_C(superfamily pfam02837:Glyco_hydro_2_N(superfamily pfam12690:BsuPI(superfamily:cl15037

Related clusters [12]

Cluster	Name	Distance	Protein	Median length (aa)	Genomes
PCLA_3326696	beta-galactosidase	0.488	24	1037	23
PCLA_2551175	beta-galactosidase	0.69	9	1037	9
PCLA_875628	beta-galactosidase	0.69	9	1037	9
PCLA_2547513	beta-D-galactosidase subunit A	0.813	13	1034	13
PCLA_854227	glycoside hydrolase	0.816	31	1013	30
PCLA_377727	beta-galactosidase	0.817	32	626	29
PCLA_2757734	beta-galactosidase				
PCLA_823497	beta-galactosidase				
PCLA_5003939	beta-galactosidase				
PCLA_912517	beta-galactosidase				

Genome groups (clades)

Clade ID	Name	Proteins in Cluster	Total Annotated Genomes	Proteins per Genome (median)
492	Escherichia coli	59	486	5156
508	Yersinia pestis	16	88	4364
495	Enterobacter cloacae complex			
496	Klebsiella pneumoniae			
504	Pectobacterium			
505	Dickeya	4	4	4144

Filters

Search by organism name, locus tag or protein name

escherichia

Hide identical proteins:

Selected by escherichia

Protein Table

Clade ID	Organism	Protein name	Accession	Locus_tag	Length (aa)	UniProtKB / SwissProt	Identical group	BLINK
444	Aeromonas hydrophila subsp. hydrophila ATCC 7966	beta-D-galactosidase	YP_858525	AHA_4101	1025	BGAL_AERHH	WP_011707763	◆
443	Aeromonas veronii B565	glycoside hydrolase family 2 TIM barrel	YP_004390763	B565_0111	1019			
2065	Citrobacter koseri ATCC BAA-895	beta-D-galactosidase	YP_001454367	CKO_02825	1025	BGAL_CROS8	WP_012125561	◆
494	Citrobacter rodentium ICC168	beta-galactosidase	YP_003364037	ROD_03991	1027			
502	Cronobacter sakazakii ATCC BAA-894	beta-D-galactosidase	YP_001439042	ESA_02977	1043	BGAL_CROS8	WP_012125561	◆
502	Cronobacter sakazakii ES15	beta-D-galactosidase	YP_006344151	ES15_3067	1084		WP_014729501	◆

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.

beta-D-galactosidase [Escherichia coli str. K-12 substr. MG1655] (A)

NCBI Reference Sequence: NP_414878.1
GenPept FASTA

Link To This Page | Feedback (D)

Protein Features
beta-D-galactosidase

Mature Peptide AA Features
beta-D-galactosidase

Region Features - CDD
lacZ

Glyco_hydro_2_N Glyco_hydro_2_C

beta-D-galactosidase
Mat-Peptide AA: beta-D-galactosidase
Location: 2..1,024
Length: 1,023 (B)

Analyze this sequence (D)
Run BLAST
Identify Conserved Domains
Highlight Sequence Features

Protein 3D Structure
E. Coli (lacZ) Beta-galactosidase (g974a)
2-deoxy-galactosyl-
PDB: 4DUV
Source: Escherichia coli K-12
Method: X-Ray Diffraction
Resolution: 2.1 Å
See all 45 structures...

Articles about the lacZ gene (E)

Protein clusters for NP_414878.1 (F)
Beta-D-galactosidase - forms a homotetramer; hydrolyzes lactose disaccharide to galactose and glucose;
Total proteins: 153
Total genera: 19
Conserved in: Bacteria

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.

Conserved domains on [gi|16128329|ref|NP_414878|]
beta-D-galactosidase [Escherichia coli str. K-12 substr. MG1655]

View full result (F)

Graphical summary show options ▶

Query seq. 1 125 250 375 500 625 750 875 1024

Specific hits
Superfamilies
Multi-domains

Glyco_hydro_2_N Glyco_hydro_2_C Bgal_small_N
Glyco_hydro_2_N superfam1 Glyco_hydro_2 su Glyco_hydro_2_C superfamily Bgal_small_N superfamily

lacZ

Search for similar domain architectures (H) Refine search (H)

List of domain hits

Description	PssmId	Multi-dom	E-value
[+]Glyco_hydro_2_C[pfam02836], Glycosyl hydrolases family 2, TIM barrel domain; This family contains beta-galactosidase, beta-mannosidase ar	217247	no	1.19e-148
[+]Bgal_small_N[pfam02929], Beta galactosidase small chain; This domain comprises the small chain of dimeric beta-galactosidases EC:3.2.1.111779	111779	no	3.25e-115
[+]Glyco_hydro_2_N[pfam02837], Glycosyl hydrolases family 2, sugar binding domain; This family contains beta-galactosidase, beta-mannosidas	217248	no	1.30e-59
[+]Glyco_hydro_2[pfam00703], Glycosyl hydrolases family 2; This family contains beta-galactosidase, beta-mannosidase and beta-glucuronidase	216070	no	2.06e-14
[+]lacZ[PRK09525], beta-D-galactosidase: Reviewed	236548	yes	0e+00

Cd Length: 1027 Bit Score: 2151.96 E-value: 0e+00

gi 16128329 1 MIMIDSLAVLQRRDWNPGVTQLNRLAAHPPFASWRNSEEARIDRPSQQLRSLNGEWRFAWFPAPPAEAVPESWLECDLP 80
Cdd: PRK09525 1 MIMIDSLAQILARRDWNPGVTQLNRLPAHPPFASWRNSEEARIDRPSQQRQSLNGEWRFSYFPAPPAEAVPESWLECDLP 80

gi 16128329 81 EADITVVVPSNWLQMHGYDAPITYNTVYPIFVNPFFVPTENPTGCYSLTFNVDESWLQEGQTRIIIFDGVNSAFHLWCNGRNV 160
Cdd: PRK09525 81 DADITFVPSNWLQMHGYDAPITYNTVYPIFVNPFFVPTENPTGCYSLTFTVDESWLQSGQTRIIIFDGVNSAFHLWCNGRNV 160

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.

Results of BLAST (show results in standard format) (G)

Query: gj|530777049 glycoside hydrolase family2 [uncultured bacterium] Length: 1023aa (C)

803 hit proteins are represented by 618 proteins

Genus-level clusters are represented with a plus sign that can be expanded to see other proteins from that cluster (BLAST results not available for (D) proteins). Both organism and score are sortable. The organism, protein name, accession, and locus_tag are links to taxonomy, protein, and gene Entrez database.

Organism	Protein Name	Accession	Length	Locus tag	Cluster	Blink	BI2Seq	Score (Bits)	E-Value
Bifidobacterium longum subsp. longum JDM301	family 2 glycoside hydrolase	YP_003661049	1023aa	BLJ_0749	PCLA_2502852	◆	◆	1925.13	0
Bifidobacterium adolescentis ATCC 15703	beta-galactosidase	YP_910468	1023aa	BAD_1605	PCLA_2502852	◆	◆	1632.4	0
Bifidobacterium adolescentis ATCC 15703	beta-galactosidase	YP_910445	1049aa	BAD_1582	PCLA_2502852	◆	◆	1359.3	0
Bifidobacterium longum subsp. longum									
BBMN68 (represents 2 sequences)	lacZ1	YP_004001410	1063aa	BBMN68_1812	PCLA_2502852	◆	◆	1122.31	0 (E)
Gardinerella vaginalis ATCC 14019 (represents 2 sequences)	beta-galactosidase	YP_003985209	1050aa	HMPREF0421_20100	PCLA_2502852	◆	◆	1074.63	0
Bifidobacterium dentium Bd1	beta-galactosidase	YP_003361063	975aa	BDP_1647	PCLA_2502852	◆	◆	1070.43	0
Bifidobacterium animalis subsp. lactis AD011 (represents 9 sequences)	beta-galactosidase	YP_002469329	1067aa	BLA_0454	PCLA_2502852	◆	◆	1024.85	0
Bifidobacterium bifidum S17 (represents 3 sequences)	beta-galactosidase	YP_003939185	1052aa	BBIF_1406	PCLA_2502852	◆	◆	1023.8	0
Bifidobacterium bifidum BGN4	beta-galactosidase	YP_006394909	1052aa	BBB_1439	PCLA_2502852	◆	◆		
Bifidobacterium bifidum PRL2010	beta-galactosidase	YP_003971530	1052aa	BBPR_1460	PCLA_2502852	◆	◆		
Bifidobacterium bifidum PRL2010	beta-galactosidase	YP_003971530	1052aa	BBPR_1460	PCLA_2502852	◆	◆	1022.75	0 (F)

Reference, help documentation and FTP

- [1] The National Center for Biotechnology Information's Protein Clusters Database. Klimke W, et al. 2009. Nucleic Acids Res. 37: D216-23 (www.ncbi.nlm.nih.gov/pubmed/18940865).
- [2] NCBI Prokaryotic Genomes Automatic Annotation Pipeline: www.ncbi.nlm.nih.gov/genomes/static/Pipeline.html
- [3] ProtClustDB Help Manual in NCBI Bookshelf: www.ncbi.nlm.nih.gov/books/NBK3797/
- [4] Data dump for protein cluster release through FTP: <ftp://ncbi.nih.gov/genomes/CLUSTERS/>