NCBI Entrez Protein Clusters

A collection of Proteins Grouped by Sequence Similarity and Fuction

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

In order to facilitate rapid protein level analysis and annotation, the National Center for Biotechnology Information (NCBI) prokaryotic Reference Sequence (RefSeq) group has created a new prokaryotic protein cluster database.

The NCBI Entrez Protein Clusters database is a collection of Reference Sequence (RefSeq) proteins from the complete genomes of prokaryotes, plasmids, and organelles grouped and annotated based on sequence similarity and protein function. Proteins are automatically grouped into clusters based on reciprocal best-hit BLAST scores.

www.ncbi.nlm.nih.gov/sites/entrez?db=proteinclusters

Overview

The clusters are divided into curated and non-curated sets. Curated clusters have consistent nomenclature, protein function descriptions, and publication information that is displayed on the overview page along with metabolic functions, conserved domains, and several tools for displaying the proteins (see Figure 1). Non-curated clusters display the same automatically collected information.

The protein cluster overview page also provides a table showing each protein in the cluster, the organism from which it was derived, links to Entrez Protein and Gene, pre-computed BLAST results in Blink, upstream and downstream clusters in the local genomic neighborhood (color-coded according to COG functional group), and alignment and domain information (color-coded to differentiate each domain) (shown in Figure 2).

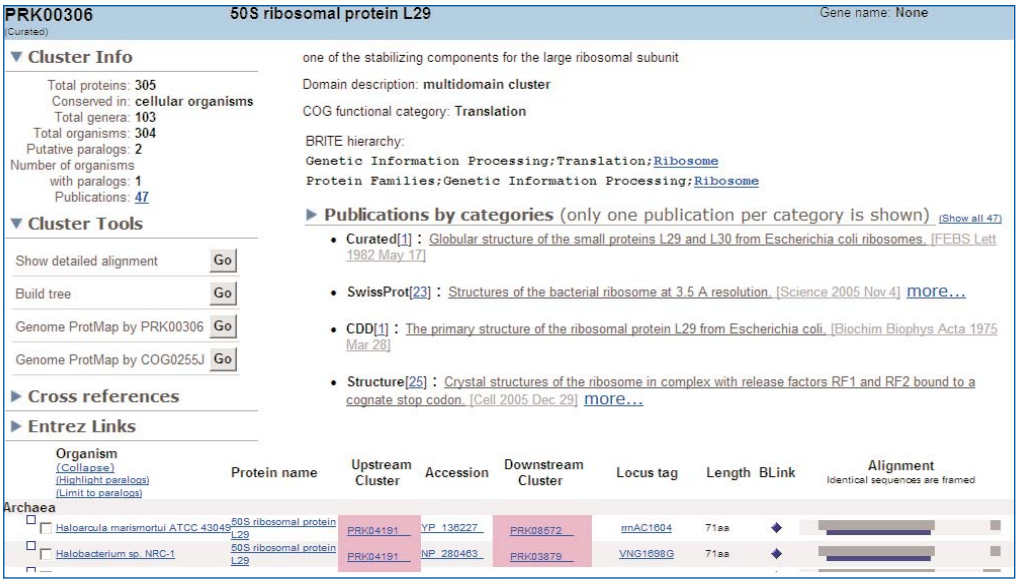


Figure 1. Overview page for 50S ribosomal protein L29.

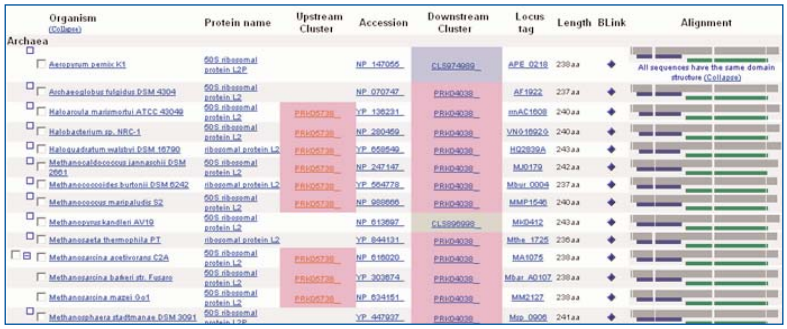


Figure 2. Protein table for 50S ribosomal protein L2.

Tools

Tools are provided to display detailed alignment, phylogenetic relalationships or genome neighborhood of the proteins in the cluster.

The detailed alignment displays a pre-computed multiple sequence alignment with domain information, protein sequence features (via structures), and amino acid display (either by similarity or by consensus).

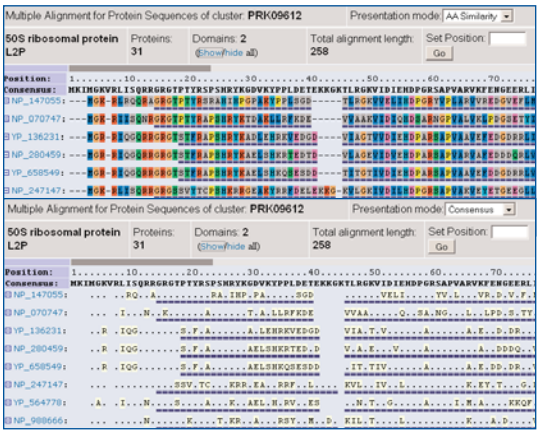


Figure 3. Detailed alignment with expanded domain information for 50S ribosomal protein L2. On the top the amino acids are displayed by similarity while on the bottom they are shown by consensus.

The interactive phylogenetic tree allows different distance and tree construction methods, varying levels of display, and collapse/squeeze, highlighting and re-rooting functions.

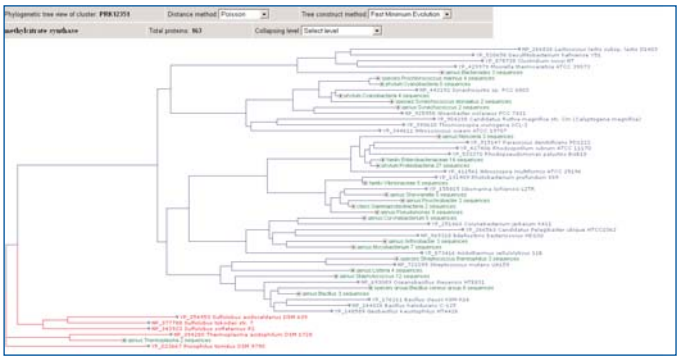


Figure 4. Phylogenetic tree display for the methylcitrate synthase cluster with archaeal proteins highlighted in red via the protein table on the overview page.

ProtMap shows the local genomic neighborhood for either a protein cluster or for a Cluster of Orthologous Group. The genes encoding the cluster in question are highlighted in the middle of the screen while adjacent genes are color-coded according to the COG functional group to which they belong.

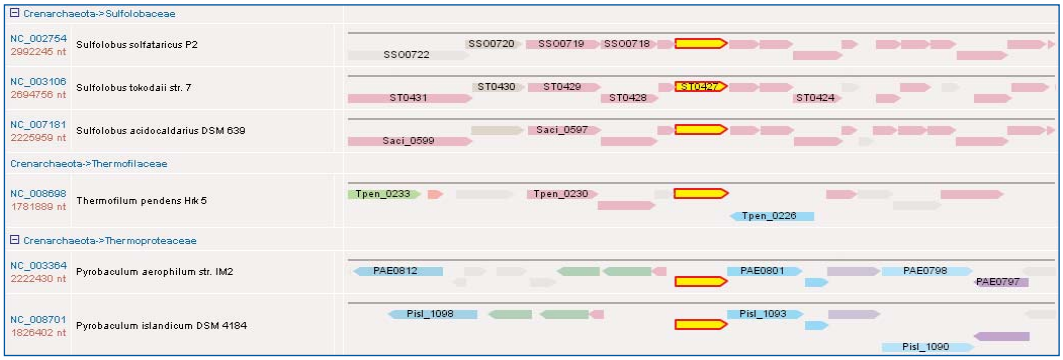
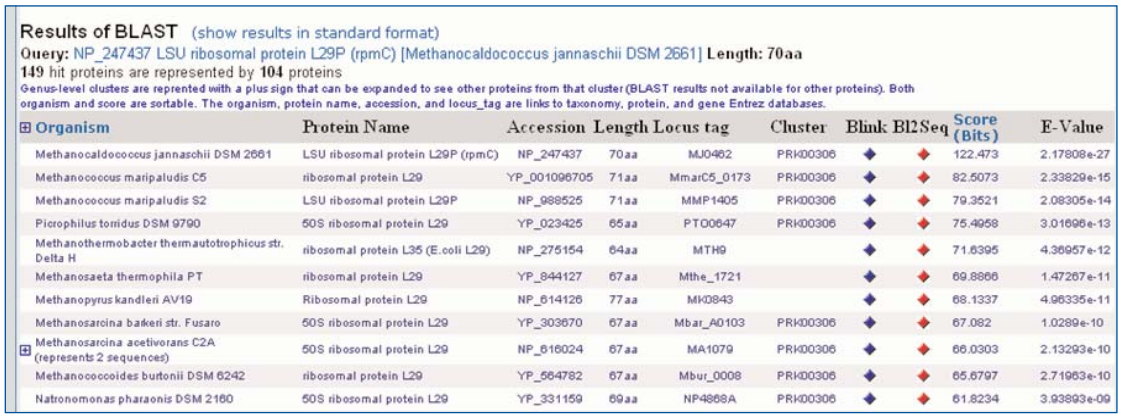


Figure 5. Protmap display for ribosomal protein L2 centered on archaeal proteins. In the Sulfolobaceae, this gene is surrounded by other ribosomal genes (purple shaded genes), while in the Thermoproteaceae it is not (blue shaded genes).

Search

There are numerous ways to query protein clusters, either with search terms in Entrez (Protein Clusters uses all of the normal Entrez features) or with a protein or nucleotide sequence.

There are two ways to query protein clusters using sequences: Concise Protein BLAST and CD-Search. The Concise Protein BLAST database can be queried with either a protein or nucleotide sequence. This database consists of both curated and non-curated prokaryotic protein clusters and each cluster has been sliced at the level of genera to provide "subclusters" from which a single protein sequence representative has been chosen (randomly).



Results of BLAST (show results in standard format)

Query: NP_247437 LSU ribosomal protein L29P (rpmC) [Methanocaldococcus jannaschii DSM 2661] **Length: 70aa**

149 hit proteins are represented by **104** proteins

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

Organism	Protein Name	Accession	Length	Locus tag	Cluster	Blink	Bl2Seq	Score (Bits)	E-Value
Methanocaldococcus jannaschii DSM 2661	LSU ribosomal protein L29P (rpmC)	NP_247437	70aa	MJ0462	PRK00306			122.473	2.17808e-27
Methanococcus maripaludis C5	ribosomal protein L29	YP_001096705	71aa	MmarC5_0173	PRK00306			82.5073	2.33829e-15
Methanococcus maripaludis S2	LSU ribosomal protein L29P	NP_988525	71aa	MMP1405	PRK00306			79.3521	2.08305e-14
Picrophilus torridus DSM 9790	50S ribosomal protein L29	YP_023425	65aa	PTO0647	PRK00306			75.4958	3.01696e-13
Methanothermobacter thermautotrophicus str. Delta H	ribosomal protein L35 (E.coli L29)	NP_275154	64aa	MTH9				71.6395	4.36957e-12
Methanosaeta thermophila PT	ribosomal protein L29	YP_844127	67aa	Mthe_1721				69.8866	1.47267e-11
Methanopyrus kandleri AV19	Ribosomal protein L29	NP_614126	77aa	MK0843				68.1337	4.96335e-11
Methanosarcina barkeri str. Fusaro	50S ribosomal protein L29	YP_303670	67aa	Mbar_A0103	PRK00306			67.082	1.0289e-10
Methanosarcina acetivorans C2A (represents 2 sequences)	50S ribosomal protein L29	NP_616024	67aa	MA1079	PRK00306			66.0303	2.13293e-10
Methanococcoides burtonii DSM 6242	ribosomal protein L29	YP_564782	67aa	Mbur_0008	PRK00306			65.6797	2.71963e-10
Natronomonas pharaonis DSM 2160	50S ribosomal protein L29	YP_331159	69aa	NP4868A	PRK00306			61.8234	3.93893e-09

Figure 6. Concise Protein BLAST results for protein NP_247427 LSU ribosomal protein L29P. The result for *Methanosarcina acetivorans* represents both *M. acetivorans* and *M. mazei.*

CD-Search can only be queried using a protein sequence. Position-specific scoring matrices (PSSMs) are constructed from the alignments of each of the curated clusters. The PSSMs for curated prokaryotic clusters have been added to the CD-Search page from the Conserved Domain database and can be searched using rps-blast.

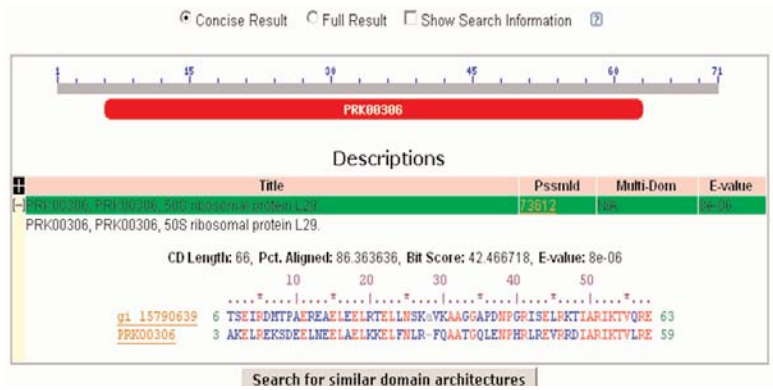


Figure 7. CD-Search results for protein NP_247427 LSU ribosomal protein L29P.