

Protein Clusters

Tatiana Tatusova, PhD,¹ Leonid Zaslavsky, PhD,¹ Boris Fedorov, PhD,¹ Diana Haddad, PhD,¹ Anjana Vatsan, PhD,¹ Danso Ako-adjei, PhD,¹ Olga Blinkova, PhD,¹ and Hassan Ghazal, PhD^{1,2}

Created: September 14, 2014.

Scope

The Protein Clusters dataset consists of organized groups (clusters) of proteins encoded by complete and draft genomes from the NCBI Reference Sequence (RefSeq) collection of microorganisms: prokaryotes, viruses, fungi, protozoans; it also includes protein clusters from RefSeq genomes of plants, chloroplasts, and mitochondria. Clusters for each group are created and curated separately and given a different accession prefix. The primary goal of Protein Clusters is to provide the support to functional annotation of RefSeq genomes. Functional annotation of novel proteins is based on the assumption that proteins with high levels of sequence similarity are likely to share the same function. This oversimplified model of a linear evolution where similar proteins evolve from a single ancestor is further complicated by the events of gene duplication. Clusters of related (homologous) proteins include both orthologs and paralogs. Orthologs are genes in different organisms (species) that evolved from a common ancestral gene by speciation; paralogs are genes related by duplication within a genome. The definition was first introduced by Fitch in 1970 (7). The analysis of protein families from various organisms has shown that this definition does not embrace all the complexity of relationships of genes from different organisms. For more details see Koonin *et al.* 2005 (16). The NCBI Protein Clusters dataset contains automatically generated clusters that do not distinguish orthologs and paralogs. During manual evaluation some clusters containing paralogs can be split by curators, especially if the paralogs are known to have different functions.

History

The first complete bacterial genome of *Haemophilus influenzae* Rd KW20 sequenced and released in 1995 opened a new era in genome analysis (8). In the following year four more genomes were completed producing an unprecedented variety of protein sequences from all three major kingdoms (Archaea, Eubacteria, and Eukaryota). Comparative analysis of homologous genes has been used in evolutionary studies and functional classification since the first sequence became available, but for the first time the whole proteome of several organisms became available for comparison. New genome-scale methods were needed to provide an understanding of the true orthologous relationships of protein sequences. The protein database of Clusters of Orthologous Groups (COGs), a pioneering work of NCBI scientists, was the first attempt at creating a phylogenetic classification of the complete complement of proteins encoded by complete genomes (23). The COG approach is based on the simple notion that any group of at least three proteins from distant genomes that are more similar to each other than they are to any other proteins from the same genome are most likely to form an orthologous set. The COG project has proved to be an excellent approach for understanding microbial

evolution and the diversity of protein functions encoded by their genomes. However, the major difficulty of any genomic data resource in the modern era of rapid genomic sequencing is keeping the genomic data and the annotations up-to-date.

The RefSeq project, which contains non-redundant sets of curated transcript, gene, and protein information in eukaryotes, and gene and protein information in prokaryotes, has been a very successful way to maintain and update annotated data. Given the increasing number of prokaryotic genomes being deposited, it became apparent that annotating protein families as a group was a convenient and efficient way to functionally annotate these data. The Protein Clusters database was constructed with two goals in mind: first, to routinely update RefSeq genomes with curated gene and protein information from such clusters; and second, to provide a central aggregation source for information collected from a wide variety of sources that would be useful for scientists studying protein-level or genomic-level molecular functions. In addition, curators routinely parse the scientific literature for reports of experimentally verified functions as the basis for existing or potential connections to genes/proteins, and such connections are added as annotations on each cluster. The first release of NCBI Protein Clusters in 2007 contained ~1 million proteins encoded by complete chromosomes and plasmids from three major groups: prokaryotes, bacteriophages, and the organelles (15). Since then the scope has been expanded to other taxonomic groups and proteins from draft genomes.

As of April 2013 the dataset represents more than 20 million proteins.

Data Model

Clustering is a well-known method in statistics and computer science. For a given set of entities, clusters are defined as subsets that are homogeneous and well separated. The cluster analysis should start from a definition of homogeneity and separation. Most clustering methods rely upon similarities (or distance) between entities. Protein clusters are aimed to be groups of homologous proteins. The similarity between two protein sequences is measured by maximum alignment between the sequences calculated by BLAST. There are multiple ways of defining various types of clusters that are based on criteria used to express separation or homogeneity of a cluster and separation from other clusters. NCBI Protein Clusters uses two methods for clustering, both resulting in building cliques, one based on partitioning and the other based on hierarchical aggregation.

Once clustered, each protein cluster is assigned a cluster ID and accession (letter prefix followed by digits) that is stable from version to version as long as the majority of its proteins don't change. A protein cluster also includes certain attributes aggregated from proteins: "Gene names" (locus), "COG functional categories," "EC numbers," and "Conserved Domains." An attribute "Conserved in" defines the common taxonomical name of genomes included in the cluster. The Protein Clusters database also includes a set of "Related Clusters". Besides these attributes, each protein record in the database has "Organism name," "Protein name," "Protein accession," "Locus tag," "Length," and UniProtKB / SwissProt accession. These attributes are easily searchable within a cluster and also through the whole database.

Statistics

Proteins:	31
Conserved in:	Bacteria
Total genera:	7
Total organisms:	28
Putative Paralogs:	3
Loci:	sacC, sacC1
COG functional category:	Carbohydrate transport and metabolism

Related Clusters

cluster	Name	Distance	Protein	Median length(aa)	Genomes
PCLA_776568	glycosyl hydrolase family 32	0.529	51	516	47
PCLA_5026923	fructan hydrolase	0.544	2	507	2
PCLA_5501389	levanase	0.569	2	446	2
PCLA_5502029	Levanase	0.578	2	509	2
PCLA_376321	levanase	0.587	21	677	21
PCLA_728286	glycosyl hydrolase family 32	0.601	19	509	13
PCLA_2508750	glycosyl hydrolase family 32	0.608	2	485	1
PCLA_2993089	levanase	0.61	3	635	3
PCLA_3254773	levanase	0.614	6	492	6
PCLA_5838420	beta-fructosidase, levanase/invertase	0.615	2	497	2

Genome Groups (clades)

Clade ID	Name	Proteins in Cluster	Total Annotated Genomes	Proteins per Genome (median)
19970	Paenibacillus	5	9	5268
19976	Paenibacillus mucilaginosus	3	3	7330
19973	Paenibacillus	2	3	6237
19975	Paenibacillus elgii	2	1	7776
21852	Paenibacillus sp. A9	2	1	4856
21853	Paenibacillus sp. PAMC 26794	1	1	5873

Protein Table

Clade ID	Organism	Protein name	Accession	Locus_tag	Length	Identical group	BLINK
22152	Actinopolyspora mortivallis DSM 44261	hypothetical protein	WP_019853821	ACTMO_06285	514	WP_019853821	◆
21263	Amphibacillus jilinensis Y1	hypothetical protein	WP_017470882	B494_02995	484	WP_017470882	◆
21267	Bacillus alcalophilus ATCC 27647	glycosyl hydrolase family protein	WP_003322074	BalcAV_07657	477	WP_003322074	◆
21936	Bacillus bataviensis LMG 21833	SacC2	WP_007084638	BABA_08076	492	WP_007084638	◆
21271	Bacillus endophyticus 2102	hypothetical protein	WP_019393615	A360_15930	531	WP_019393615	◆

Example of a bacterial cluster PCLA_5029913 glycoside hydrolase

Clustering Methods

Partitioning in Cliques

Proteins are compared by sequence similarity using BLAST all against all (E-value cutoff $10E-05$; effective length of the search space is set to $5 \times 10E8$). Each BLAST score is then modified by protein length \times alignment length of the BLAST hit and the modified scores are sorted. Clusters (also known as cliques) consist of protein sets such that every member of the cluster hits every other protein member (reciprocal best hits by modified score). Cluster membership is such that for any given protein in the cluster (protein A), all the other members of the cluster will have a greater modified score to protein A than any protein outside of that cluster. During clustering, there are no cutoffs used nor strict requirements for clusters of orthologous groups, nor any check on phylogenetic distance. The initial set of uncurated clusters created in 2005 was used as a starting point for

curation and has been updated periodically since then. During updates, new proteins are added to curated clusters. In the uncurated cluster set, proteins are allowed to repartition into different cluster sets, although this happens rarely and usually only in the case of smaller clusters.

Hierarchical Aggregation in Cliques

A new approach implemented for prokaryotic genomes is based on hierarchical clustering. While a hierarchical structure is conventionally represented by a dendrogram and clusters are selected as a sub-tree corresponding to a certain threshold (14, 17, 18), the hierarchical structure goes beyond simple clustering (1, 3). First, all the proteins are organized in global clusters, then links between clusters are calculated reflecting the similarity between the clusters based on several criteria.

Protein Clustering Procedure

The similarity of proteins is determined from the aggregated BLAST hits obtained by *blastp* with e-value 10^{-3} . Two proteins are considered connected if there is an aggregated BLAST hit between them satisfying criteria on hit length and score. More specifically, we require the aggregated hit lengths on each protein, $l_{ij}^{(1)}$ and $l_{ij}^{(2)}$, satisfy the inequalities $l_{ij}^{(1)} \geq \varepsilon \cdot l_i$ and $l_{ij}^{(2)} \geq \varepsilon \cdot l_j$, where l_i and l_j are lengths of proteins, and $0.5 < \varepsilon < 1$, and the aggregated BLAST score S_{ij} satisfy the inequality $S_{ij} \geq \gamma \cdot \max(S_{ii}, S_{jj})$, where S_{ii} and S_{jj} are self-scores.

The modified BLAST distance is defined as

$$d_{ij}^{\alpha} = 1 - \frac{S_{ij}}{\max(\chi_{ij}^{(1)} \cdot S_{ii}, \chi_{ij}^{(2)} \cdot S_{jj}, S_{ij})},$$

where the score modifications are $\chi_{ij}^{(1)} = \max\left(\frac{l_{ij}^{(1)}}{l_i}, 1-\alpha\right)$ and $\chi_{ij}^{(2)} = \max\left(\frac{l_{ij}^{(2)}}{l_j}, 1-\alpha\right)$, and $0 < \alpha \ll 1$. Using $\alpha > 0$

allows some flexibility at the end of the sequences (the distance is reduced to $d_{ij}^0 = 1 - \frac{S_{ij}}{\max(S_{ii}, S_{jj})}$ when $\alpha = 0$).

Clusters are aggregated in a hierarchical manner using the complete linkage distance, with an additional requirement that the minimum distance between clusters $d^{\min}(\Lambda, \Omega)$ should not exceed threshold δ , where $0 < \delta < 1 - \gamma$, for clusters Λ and Ω to be merged. Note that we calculate and use both $d^{\min}(\Lambda, \Omega)$ and $d^{\max}(\Lambda, \Omega)$ in our clustering procedure (see Figure 1). Because of the sparse nature of connections and applied thresholds, we build a family of trees that we consider clusters. Currently, we use the values $\varepsilon = 0.7$, $\gamma = 0.2$, $\alpha = 0.1$, and $\delta = 0.5$.

Establishing Links between Related Clusters

Each protein within a cluster should be similar to *all* other proteins in the same cluster, satisfying coverage and similarity criteria. Still, a pair of proteins in different clusters could be similar. Such clusters are designated as *related clusters* (1, 3, 12, 24). Links between related clusters are stored in *link indexes*, which are used to show neighborhoods of clusters in Entrez search.

Organization of computations. First, we eliminate redundancy and near-redundancy in the protein dataset (2, 12). Representative proteins from groups of redundant and nearly-redundant proteins are selected by the program USEARCH (5).

In order to perform clustering in parallel, the dataset is partitioned in disjoint sets (Figure 2) using a parallel implementation based on a disjoint-set forest with union-by-rank heuristics (4, 22), and then clustering is performed concurrently in partitions. When looking for disjoint sets, we only consider connections with $d_{ij}^{\alpha} \leq \delta$.

After the clustering is performed, link indexes are also calculated in parallel from the aggregated BLAST hit and protein assignment to clusters.

Dataflow

Input data are proteins from complete and draft (WGS) genomes that pass certain quality filters.

Proteins marked as incomplete in metadata (“incomplete,” “no start,” “no end,” “fragment,” etc.) are removed and only proteins that are presumed complete are analyzed. Bacterial genome clustering has a different dataflow compared to other genomes as indicated in Figure 3.

Manual Curation

One of the most important aspects of the curation process of Protein Clusters is the assignment of function that is obtained from the literature. Curated functional annotation can be propagated to all proteins within the cluster. That process improves the functional annotation of RefSeq genomes and unifies and standardizes the naming rules across various organisms and different annotation pipelines. In addition to providing functional annotation that is required for each cluster, other data are also added, such as the gene name, a detailed description about the protein, the E.C. number, and relevant publications.

Cluster Display

A protein cluster is represented by a list of protein identifiers (accessions) and the genomes that code for the proteins. Each cluster has a stable unique identifier and a functional cluster name. The cluster name is automatically calculated and followed by manual review. Each cluster provides statistics to indicate the number of proteins within that cluster and other important features including the Protein Table.

Cluster Examples

Viruses

The ease and efficiency of nucleic acid sequencing has led to an abundance of sequence data. Because of the relatively small genome size of viruses, the influx of sequence data has been particularly large. Likewise, the ever increasing advancements and publications in virology research make it difficult for researchers to keep up with new discoveries in protein structure and function. Rapid viral evolution, combined with the relatively large number of strains and closely related species in most viral families, makes the Protein Clusters resource an ideal channel through which viral RefSeq genomes can be curated.

The *Poxviridae* is an example of a virus family with a large set of proteins having varying degrees of similarity in function, homology, and structure (13). The poxvirus RNA helicase NPH-II belongs to a family of ubiquitous ATP-dependent helicases that are required for RNA metabolism in bacteria, eukaryotes, and many viruses (6). The NPH-II family of helicases found in hepatitis C and various poxviruses have similar sequence, structure, and mechanisms of action that are essential for viral replication. The protein cluster PHA2653 includes 27 NPH-II helicase proteins from various members of the *Poxviridae*. While they share a high level of homology, evolutionary pressures have resulted in changes to both sequence and activity. Of particular interest is the fact that the poxvirus NPH-II belongs to a superfamily, SF2, of which several eukaryotic helicases that play a major role in cellular responses to viral infection also belong (19). Furthermore, the helicase core domain is a component of the dicer complex which mediates RNAi in higher eukaryotes (9). Therefore, it stands to reason that study of the NPH-II helicases of the *Poxviridae* can serve as a model for understanding several distinct biological processes.

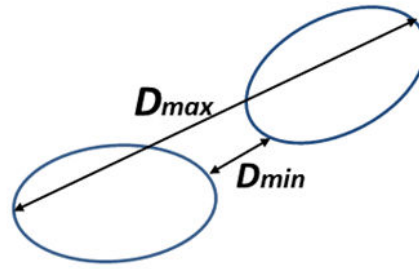


Figure 1. Minimal and maximum distance between clusters

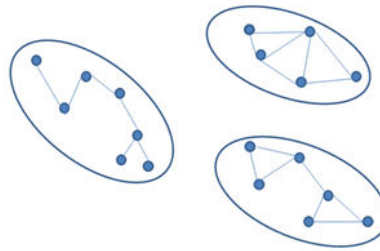


Figure 2. Disjoint sets.

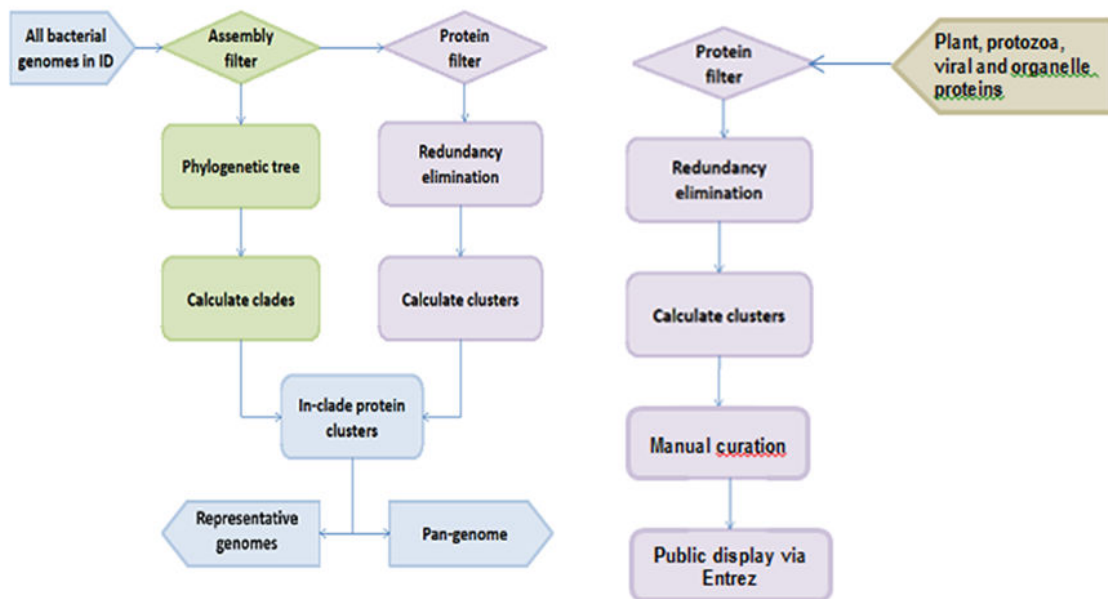


Figure 3. Dataflow for prokaryotic and eukaryotic genomes

Frequently, several alternative names are used for viral proteins; this variation can lead to confusion for researchers and slow scientific progress. To standardize protein names, NCBI staff (viral genome curators) work closely with viral protein experts from UniProt. Such collaborations have resulted in functional naming for viral protein clusters from the family *Adenoviridae*. One of their representatives is cluster PHA3614. It combines related and highly conserved proteins from the genus *Mastadenovirus*, which presumably play an important role in host modulation (11). The old, commonly used name of proteins from the PHA3614 cluster was the E3 12.5

kDa protein. Because the size of the protein could vary in different viruses without its biological role changing, the molecular mass, as a component of protein name, was not informative and could be misleading. Therefore we proposed a new, functional name for this cluster: “putative host modulation protein E3.” All existing synonyms were included in the cluster’s functional description. Since the existence of the putative host modulation protein E3 was experimentally supported only for human adenovirus 2 and human adenovirus 5, this information was also included in the description of the cluster. These changes will be visible with the next cluster update.

Protozoa

The following is an example of the significance of publication links in a cluster of proteins as a tool to identify orthologs and paralogs.

PTZ00021 falcipain-2 ID: 2458473

Cysteine proteases identified and characterized in *Plasmodium falciparum*. Hydrolyses the erythrocyte hemoglobin into its amino acid constituents, which are used by the parasite for protein synthesis.

Statistics

Proteins: 9
 Conserved in: **Plasmodium**
 Total genera: 1
 Total organisms: 3
 Putative Paralogs: 0
 Locus:
 Structures: 6
 CDDs: [PTZ00021](#), [smart00848:Inhibitor_I29\(superfamily:cl07031\)](#), [pfam08246:Inhibitor_I29\(superfamily:cl07031\)](#), [cd02248:Peptidase_C1A\(superfamily:cl00298\)](#)

Filters

Protein Table

Organism	Protein name	Accession	Locus_tag	Length (aa)	BLINK
<i>Plasmodium falciparum</i> 3D7	falcipain-2B	XP_001347832	PF11_0161	482	◆
<i>Plasmodium falciparum</i> 3D7	falcipain-3	XP_001347833	PF11_0162	492	◆
<i>Plasmodium falciparum</i> 3D7	falcipain-2A	XP_001347836	PF11_0165	484	◆
<i>Plasmodium knowlesi</i> strain H	<i>P. knowlesi</i> ortholog of falcipain	XP_002259151	PKH_091240	477	◆
<i>Plasmodium knowlesi</i> strain H	<i>P. knowlesi</i> ortholog of falcipain	XP_002259152	PKH_091250	495	◆
<i>Plasmodium knowlesi</i> strain H	<i>P. knowlesi</i> ortholog of falcipain	XP_002259153	PKH_091260	479	◆
<i>Plasmodium vivax</i> Sal-1	vivapain-2	XP_001615272	PVX_091405	484	◆
<i>Plasmodium vivax</i> Sal-1	vivapain-3	XP_001615273	PVX_091410	495	◆
<i>Plasmodium vivax</i> Sal-1	vivapain-2	XP_001615274	PVX_091415	487	◆

This is a cysteine protease, originally identified and characterized in *Plasmodium falciparum*; it hydrolyses the erythrocyte hemoglobin into its amino acid constituents, which are used by the parasite for protein synthesis. In this cluster of proteins, “falcipain” is present in 4 different *Plasmodium* species. Falcipain-2 differs from the falcipains in other species (i.e., vivapain -2 and -3, berghepain, etc.) as well as within the *P. falciparum* (i.e., falcipain -3) in sequence, in the timing of expression and in the acidic environment needed for enzymatic activation, but they all appear to have the same function (20). Interestingly, the two *P. falciparum* falcipain-2 proteins in this cluster are each located in a different part of chromosome 11, although they share high amino acid sequence homology and a seemingly identical function. The differences here also appear to be in expression timing and in the level of expression. Falcipain-2A (PF11_0165) appears to be expressed earlier in the trophozoite stage and in higher amounts than falcipain-2B (PF-11_0161) (10, 20).

Also of interest is the fact that cysteine protease inhibitors have been shown to have potent anti-malarial effects. Indeed, because this family of proteases shares low sequence identity with their human counterparts, they have been given serious consideration as potential drug targets.

Plants

Although protein clustering is not specifically geared towards clustering for orthologs or paralogs, clustering does provide a view into how different proteins are related as seen in the cluster PLN03595 shown below.

Organism	Protein name	Accession	Locus_tag	Length (aa)	UniProtKB / SwissProt
<i>Arabidopsis lyrata</i> subsp. <i>lyrata</i>	hypothetical protein	XP_002868009	ARALYDRAFT_914800	1116	
<i>Arabidopsis lyrata</i> subsp. <i>lyrata</i>	phytochrome D	XP_002868148	ARALYDRAFT_915133	1165	
<i>Arabidopsis lyrata</i> subsp. <i>lyrata</i>	hypothetical protein	XP_002868441	ARALYDRAFT_493637	1112	
<i>Arabidopsis lyrata</i> subsp. <i>lyrata</i>	phytochrome B	XP_002868263	ARALYDRAFT_480851	1163	
<i>Arabidopsis lyrata</i> subsp. <i>lyrata</i>	hypothetical protein	XP_002892510	ARALYDRAFT_471053	1122	
<i>Arabidopsis thaliana</i>	phytochrome A	NP_001117256	AT1G09570	1014	P14712
<i>Arabidopsis thaliana</i>	phytochrome A	NP_172428	AT1G09570	1122	P14712
<i>Arabidopsis thaliana</i>	phytochrome B	NP_179469	AT2G18790	1172	P14713
<i>Arabidopsis thaliana</i>	phytochrome D	NP_193360	AT4G16250	1164	P42497
<i>Arabidopsis thaliana</i>	phytochrome E	NP_193547	AT4G18130	1112	P42496, O56Y99
<i>Arabidopsis thaliana</i>	phytochrome C	NP_198433	AT5G35840	1111	P14714
<i>Brachypodium distachyon</i>	phytochrome B-like	XP_003558068	LOC100829836	1181	
<i>Brachypodium distachyon</i>	phytochrome C-like	XP_003559446	LOC100834357	1140	
<i>Brachypodium distachyon</i>	phytochrome A type 3-like	XP_003560548	LOC100836209	1131	
<i>Glycine max</i>	phytochrome A	NP_001238206	phyA	1131	
<i>Glycine max</i>	phytochrome B-like	XP_003533157	LOC100799831	1137	
<i>Glycine max</i>	phytochrome E-like	XP_003535030	LOC100808192	1120	
<i>Glycine max</i>	phytochrome B-like isoform 1	XP_003546314	LOC100794965	1149	
<i>Glycine max</i>	phytochrome E-like	XP_003546574	LOC100800339	1121	
<i>Glycine max</i>	phytochrome type A-like	XP_003554593	LOC100791098	1130	
<i>Glycine max</i>	phytochrome type A-like	XP_003555766	LOC100790763	1123	
<i>Selaginella moellendorffii</i>	hypothetical protein	XP_002991119	SELMODRAFT_161430	1143	
<i>Selaginella moellendorffii</i>	hypothetical protein	XP_002991641	SELMODRAFT_161807	1142	
<i>Sorghum bicolor</i>	hypothetical protein	XP_002463975	SORBIDRAFT_01g009930	1131	
<i>Sorghum bicolor</i>	hypothetical protein	XP_002466441	SORBIDRAFT_01g007850	1135	
<i>Sorghum bicolor</i>	hypothetical protein	XP_002467973	SORBIDRAFT_01g037340	1178	
<i>Vitis vinifera</i>	phytochrome C	XP_002268724	LOC100258014	1118	
<i>Vitis vinifera</i>	phytochrome E	XP_002271671	PHYE	1124	
<i>Vitis vinifera</i>	phytochrome B-like	XP_002278263	LOC100261882	1129	
<i>Vitis vinifera</i>	phytochrome A1	XP_002278610	PHYA	1124	

Organism	Protein name	Accession	Locus_tag	Length (aa)	UniProtKB / SwissProt
<i>Medicago truncatula</i>	Phytochrome A	XP_003591274	MTR_1g085160	1171	
<i>Medicago truncatula</i>	Phytochrome b1	XP_003594734	MTR_2g034040	1198	
<i>Medicago truncatula</i>	Phytochrome E	XP_003595711	MTR_2g049520	1122	
<i>Oryza sativa Japonica Group</i>	hypothetical protein	NP_001049910	Os03g0309200	1120	Q10MG9
<i>Oryza sativa Japonica Group</i>	hypothetical protein	NP_001051096	Os03g0719800	1128	Q10DU0
<i>Oryza sativa Japonica Group</i>	hypothetical protein	NP_001051296	Os03g0752100	1137	Q10C06
<i>Physcomitrella patens</i> subsp. <i>patens</i>	phytochrome 5c	XP_001754366	PHYPADRAFT_115388	1124	
<i>Physcomitrella patens</i> subsp. <i>patens</i>	phytochrome 5a	XP_001761145	PHYPADRAFT_206532	1123	
<i>Physcomitrella patens</i> subsp. <i>patens</i>	phytochrome 3	XP_001766035	PHYPADRAFT_185248	1123	
<i>Physcomitrella patens</i> subsp. <i>patens</i>	phytochrome 5b3	XP_001767224	PHYPADRAFT_185601	1131	
<i>Physcomitrella patens</i> subsp. <i>patens</i>	phytochrome 4	XP_001773550	PHYPADRAFT_218861	1126	
<i>Physcomitrella patens</i> subsp. <i>patens</i>	phytochrome 1	XP_001778155	PHYPADRAFT_222399	1123	
<i>Physcomitrella patens</i> subsp. <i>patens</i>	phytochrome 2	XP_001782339	PHYPADRAFT_225644	1130	
<i>Populus trichocarpa</i>	hypothetical protein	XP_002312330	POPTRDRAFT_832686	1142	
<i>Populus trichocarpa</i>	phytochrome B2	XP_002314949	POPTRDRAFT_1091155	1146	
<i>Populus trichocarpa</i>	phytochrome	XP_002318913	POPTRDRAFT_729311	1126	
<i>Ricinus communis</i>	phytochrome A, putative	XP_002512596	RCOM_1437130	1124	
<i>Ricinus communis</i>	phytochrome B, putative	XP_002519230	RCOM_1000590	1141	
<i>Ricinus communis</i>	phytochrome B, putative	XP_002519749	RCOM_0634650	1131	

PLN03595 represents a family of photoreceptors involved in the photoperiodic control of plant growth and development. This family includes diverse but structurally conserved proteins. They are expressed in different plant organs under varying light conditions. Phylogenetic analyses suggest that the phytochrome gene family is composed of four subfamilies, *PHYA*, *PHYB*, *PHYC/F*, and *PHYE*. *Arabidopsis thaliana* has an additional *PHYD*

gene that originated from the *PHYB* gene after a more recent gene duplication, and there is some functional redundancy between these two. *PHYA* and its paralog *PHYC* are found in monocots as well as in dicots, but *PHYC* is missing in some dicot lineages. Rice only has three phytochrome genes: *PHYA*, *PHYB*, and *PHYC*. Monocotyledonous plants are also known to lack several members of *PHYB* subfamily. Phytochromes exhibit distinct and cooperative functions. Mutant analysis has shown that, in rice, *phyA* and *phyB* act in a highly redundant manner to control de-etiolation under continuous red light. Under continuous far-red light, *phyA* and *phyC* are involved in photoperception, but the photoperception mode of *phyC* differs between rice and *Arabidopsis* (21).

We also used proteins of the photosynthesis system as a model for clustering validation. The photosynthesis system has been chosen as it is well conserved and characterized throughout the plant kingdom. As of now, 116 clusters were identified in plants using the “photos” keyword that were annotated and curated. The number of proteins per cluster ranged from 2 to 100. These photosynthesis proteins belong to 6 or more organisms out of 23 distinct genomes. One cluster, PLN00033, contains 22 proteins belonging to 19 organisms and corresponds to the photosystem II stability/assembly factor, which is coherent with the central role this protein plays in chloroplast biogenesis and photosystem stability (25, 26, and 27). Interestingly, 10 out of these 22 proteins from 8 different organisms are annotated as hypothetical proteins.

The second most conserved cluster, PLN00037, contains 34 proteins belonging to 18 organisms and corresponds to photosystem II oxygen-evolving enhancer protein 1 (Psb O). This situation is coherent with its crucial role in photosynthesis. Here again 11 proteins are annotated as hypothetical. Generally, the most conserved proteins in the plant kingdom are known for their central role in plant growth and development. The clustering can be used to hypothesize about the most important proteins whose function is worth analyzing further. For example, PLN03089, a cluster of 65 hypothetical proteins present in both monocots and dicots, should attract more interest. Although the proteins have homology with the Glutamate-gated kainate-type ion channel receptor subunit GluR5 in *Medicago truncatula*, there is no convincing evidence of such function.

The clusters containing protein specific to a group or subgroup of plants are also very interesting to study. Examples of such clusters are the ones with proteins present in all *viridiplantae* (PLN00046: photosystem I reaction center subunit O; PLN00054: photosystem I reaction center subunit N; PLN00049: carboxyl-terminal processing protease). The corresponding proteins would be among the most important in plant photosynthesis. Some other clusters contain proteins from a specific subgroup such as algae (PLN00100).

Access


Protein Clusters are presented in NCBI's Entrez system (<http://www.ncbi.nlm.nih.gov/sites/entrez?db=proteinclusters>)

The first public release of the Protein Clusters in NCBI's Entrez interface was in April 2007, and initially consisted of only prokaryotic clusters (15). The Entrez system provides a mechanism for the search, retrieval, and linkage between Protein Clusters and other NCBI databases, as well as external resources. Clusters can be searched by general text terms, and also by specific protein or gene names.

Limits and Advanced search allow clusters to be browsed by function and filtered by size and organism group. A table browser allows users to sort by the content of each column by clicking on the column header.

Search by organism name, locus tag or protein name

Hide identical proteins:

 Protein Table

Clade ID	Organism	Protein name	Accession	Locus_tag	Length (aa)	Identical group
22152	Actinopolyspora mortivallis DSM 44261	hypothetical protein	WP_019853821	ACTMO_06285	514	WP_019853821
21263	Amphibacillus jilinensis Y1	hypothetical protein	WP_017470882	B494_02995	484	WP_017470882
21267	Bacillus alcalophilus ATCC 27647	glycosyl hydrolase family protein	WP_003322074	BalcAV_07657	477	WP_003322074
21936	Bacillus bataviensis LMG 21833	SacC2	WP_007084638	BABA_08076	492	WP_007084638
21271	Bacillus endophyticus 2102	hypothetical protein	WP_019393615	A360_15930	531	WP_019393615
21935	Bacillus nealsonii AAU1	glycosyl hydrolase family protein	WP_016205333	A499_24244	486	WP_016205333
20034	Bacillus sp. 10403023	glycosyl hydrolase family protein	WP_010677742	B1040_010100015199	485	WP_010677742
21943	Halobacillus sp. BAB-2008	glycoside hydrolase	WP_008633984	D479_04248	533	WP_008633984
22245	Nocardiopsis salina YIM 90010	hypothetical protein	WP_017612606	D474_05620	515	WP_017612606
19975	Paenibacillus elgii B69	SacC2	WP_010492966	PelgB_010100005566	489	WP_010492966
19975	Paenibacillus elgii B69	glycoside hydrolase family protein	WP_010495814	PelgB_010100015198	492	WP_010495814
21855	Paenibacillus ginsengihumi DSM 21568	hypothetical protein	WP_019537290	F591_24885	489	WP_019537290
22129	Paenibacillus lactis 154	Glycosyl hydrolase family 32 domain protein	WP_007132088	PaelaDRAFT_4928	487	WP_007132088

Items 1 - 20 of 31 < Prev Page 1 of

Protein clusters are available for download from the FTP directory (<ftp://ftp.ncbi.nih.gov/genomes/Bacteria/CLUSTERS/>) by date and by major taxonomic groups.

Related Tools

Concise Protein BLAST

The Concise Protein database contains proteins from all clusters, as well as all singletons (not clustered proteins). From the clustered set, a representative at the genus level is chosen in order to reduce the data set. Results are therefore available rapidly and the results that are returned provide a broader taxonomic range due to this data reduction.

Concise BLAST provides an option for both protein and nucleotide searches using BLASTP and BLASTX, respectively.

<http://www.ncbi.nlm.nih.gov/genomes/prokhits.cgi>

RPS-BLAST

RPS-BLAST searches against pre-calculated position-specific scoring matrices (PSSMs) created during conserved domain processing for the CD-search tool. Therefore, only protein sequences are used for this type of search. PSSMs from the curated cluster set have been added to CDD and are also used in pre-calculated conserved domain hits available from the link menu on protein sequences and reported on each GenPept record. The curated set of PSSMs can be searched using RPS-BLAST and a protein sequence at

<http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi> or the full set of PSSMs for all curated clusters is available from FTP.

ProtMap

ProtMap is a graphical gene neighborhood tool that displays clickable, linked genes upstream and downstream of the target. The tool provides useful graphical representations of the members of a particular cluster in their genome environments. All members of the cluster of interest are mapped to their genome position, and the tool displays genomic segments coding for each member of the cluster. If the genome sequence is larger than 20KB, only the relevant 10KB portion of it is shown. Users can search for the cluster of interest by using cluster access or the COG/VOG attribute of the cluster. The display is centered on protein members of the cluster. Users can select additional sets of related proteins by clicking on the corresponding colored arrows depicting a protein, or find a cluster of interest by name, protein accession, or gene locus_tag. This resource is useful in identifying paralogs as well as missing or incorrectly annotated genes.

<http://www.ncbi.nlm.nih.gov/sutils/protmap.cgi>

References

1. Ahn YY, Bagrow J, Lehmann S. Link communities reveal multiscale complexity in networks. *Nature*. 2010 Aug 5;466(5):761–765. PubMed PMID: 20562860.
2. Cameron M, Bernstein Y, Williams HE. Clustered Sequence Representation for Fast Homology Search. *J Comput Biol*. 2007 Jun;14(5):594–614. PubMed PMID: 17683263.
3. Clauset A, Moore C, Newman ME. Hierarchical structure and the prediction of missing links in networks. *Nature*. 2008 May 1;453:98–100. PubMed PMID: 18451861.
4. Cormen TH, Leiserson CE, Rivest RL, Stein C. Introduction to algorithms. 3rd Edition, The MIT Press; 2009.
5. Edgar RC. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*. 2010 Oct 1;26(19):2460–2461. PubMed PMID: 20709691.
6. Fairman-Williams ME, Jankowsky E. Unwinding initiation by the viral RNA helicase NPH-II. *J Mol Biol*. 2012 Feb 3;415(5):819–832. PubMed PMID: 22155080.
7. Fitch WM. Distinguishing homologous from analogous proteins. *Syst Zool*. 1970 Jun;19:99–106. PubMed PMID: 5449325.
8. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*. 1995 Jul 28;269(5223):496–512. PubMed PMID: 7542800.
9. Gargantini PR, Serradell MC, Torri A, Lujan HD. Putative SF2 helicases of the early-branching eukaryote *Giardia lamblia* are involved in antigenic variation and parasite differentiation into cysts. *BMC Microbiol*. 2012 Nov 28;12:284. PubMed PMID: 23190735.
10. Goh LL, Sim TS. Homology modeling and mutagenesis analyses of *Plasmodium falciparum* falcipain 2A: implications for rational drug design. *Biochem Biophys Res Commun*. 2004 Oct 15;323(2):565–572. PubMed PMID: 15369788.
11. Hawkins LK, Wold WS. A 12,500 MW protein is coded by region E3 of adenovirus. *Virology*. 1992 Jun;188(2):486–494. PubMed PMID: 1585632.
12. Holm L, Sander C. Removing near-neighbor redundancy from large protein sequence collections. *Bioinformatics*. 1998 Jun;14(5):423–429. PubMed PMID: 9682055.
13. Hughes AL, Irausquin S, Friedman R. The evolutionary biology of poxviruses. *Infect Genet Evol*. 2010 Jan;10(1):50–59. PubMed PMID: 19833230.
14. Kaplan N, Sasson O, Inbar U, Friedlich M, Fromer M, Fleischer H, Portugaly E, Linial N, Linial M. ProtoNet 4.0: a hierarchical classification of one million protein sequences. *Nucleic Acids Res*. 2005 Jan 1;33(Database issue):D216–D218. PubMed PMID: 15608180.

15. Klimke W, Agarwala R, Badretdin A, Chetvernin S, Ciufu S, Fedorov B, Kiryutin B, O'Neill K, Resch W, Resenchuk S, Schafer S, Tolstoy I, Tatusova T. The National Center for Biotechnology Information's Protein Clusters Database. *Nucleic Acids Res.* 2009 Jan;37(Database issue):D216–23. PubMed PMID: 18940865.
16. Koonin EV. Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet.* 2005;39:309–38. PubMed PMID: 16285863.
17. Krause A, Stoye J, Vingron M. Large scale hierarchical clustering of protein sequences. *BMC Bioinformatics.* 2005 Jan 22;6:6–15. PubMed PMID: 15644130.
18. Loewenstein Y, Portugaly E, Fromer M, Linial M. Efficient algorithms for accurate hierarchical clustering of huge datasets: tackling the entire protein space. *Bioinformatics.* 2008 Jul 1;24(13):i41–49. PubMed PMID: 18586742.
19. Ranji A, Boris-Lawrie K. RNA helicases: emerging roles in viral replication and the host innate response. *RNA Biol.* 2010 Nov-Dec;7(6):775–87. PubMed PMID: 21173576.
20. Shenai BR, Sijwali PS, Singh A, Rosenthal PJ. Characterization of native and recombinant falcipain-2, a principal trophozoite cysteine protease and essential hemoglobinase of *Plasmodium falciparum*. *J Biol Chem.* 2000 Sep 15;275(37):29000–29010. PubMed PMID: 10887194.
21. Takano M, Inagaki N, Xie X, Yuzurihara N, Hihara F, Ishizuka T, Yano M, Nishimura M, Miyao A, Hirochika H, Shinomura T. Distinct and cooperative functions of phytochromes A, B, and C in the control of deetiolation and flowering in rice. *Plant Cell.* 2005 Dec;17(12):3311–3325. PubMed PMID: 16278346.
22. Tarjan RE. Data structures and network algorithms, CBMS 44, Society for Industrial and Applied Mathematics, Philadelphia, PA; 1983.
23. Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. *Science.* 1997 Oct 24;278(5338):631–637. PubMed PMID: 9381173.
24. Zaslavsky L, Tatusova T. Mining the NCBI influenza sequence database: adaptive grouping of BLAST results using precalculated neighbor indexing. *PLoS Curr.* 2009;1:RRN1124. PubMed PMID: 20029662.
25. Plücker H1, Müller B, Grohmann D, Westhoff P, Eichacker LA. The HCF136 protein is essential for assembly of the photosystem II reaction center in *Arabidopsis thaliana*. *FEBS Lett.* 2002 Dec 4;532(1-2):85–90. PubMed PMID: 12459468.
26. Meurer J, Plücker H, Kowallik KV, Westhoff P. A nuclear-encoded protein of prokaryotic origin is essential for the stability of photosystem II in *Arabidopsis thaliana*. *EMBO J.* 1998 Sep 15;17(18):5286–5297.
27. Peltier JB, Emanuelsson O, Kalume DE, Ytterberg J, Friso G, Rudella A, Liberles DA, Söderberg L, Roepstorff P, von Heijne G, van Wijk KJ. Central functions of the lumenal and peripheral thylakoid proteome of *Arabidopsis* determined by experimentation and genome-wide prediction. *Plant Cell.* 2002 Jan;14(1):211–236. PubMed PMID: 11826309.