D216–D223 *Nucleic Acids Research, 2009, Vol. 37, Database issue* *Published online 21 October 2008*
doi:10.1093/nar/gkn734

The National Center for Biotechnology Information's Protein Clusters Database

William Klimke*, Richa Agarwala, Azat Badretdin, Slava Chetvernin, Stacy Ciufo, Boris Fedorov, Boris Kiryutin, Kathleen O'Neill, Wolfgang Resch, Sergei Resenchuk, Susan Schafer, Igor Tolstoy and Tatiana Tatusova

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Department of Health and Human Services, Building 38A, 8600 Rockville Pike, Bethesda, MD 20894, USA

Received August 21, 2008; Revised September 24, 2008; Accepted October 1, 2008

ABSTRACT

Rapid increases in DNA sequencing capabilities have led to a vast increase in the data generated from prokaryotic genomic studies, which has been a boon to scientists studying micro-organism evolution and to those who wish to understand the biological underpinnings of microbial systems. The NCBI Protein Clusters Database (ProtClustDB) has been created to efficiently maintain and keep the deluge of data up to date. ProtClustDB contains both curated and uncurated clusters of proteins grouped by sequence similarity. The May 2008 release contains a total of 285 386 clusters derived from over 1.7 million proteins encoded by 3806 nt sequences from the RefSeq collection of complete chromosomes and plasmids from four major groups: prokaryotes, bacteriophages and the mitochondrial and chloroplast organelles. There are 7180 clusters containing 376 513 proteins with curated gene and protein functional annotation. PubMed identifiers and external cross references are collected for all clusters and provide additional information resources. A suite of web tools is available to explore more detailed information, such as multiple alignments, phylogenetic trees and genomic neighborhoods. ProtClustDB provides an efficient method to aggregate gene and protein annotation for researchers and is available at http://www.ncbi.nlm.nih.gov/sites/entrez?db=proteinclusters.

INTRODUCTION

From the release of the 1st complete bacterial genome for *Haemophilus influenzae* Rd KW20 in 1995 (1) to the 700th in 2008, the abundant data sets created by sequence data have been a rich resource for studying the evolution and cellular functions of a wide spectrum of microbes. One central method to understanding microbial evolution and the diversity of protein functions encoded by their genomes has been to construct protein families related by sequence similarity for analyses or distribution in publicly accessible databases. Early examples of related gene/protein family databases at NCBI include COG (Cluster of Orthologous Groups) for prokaryotes and eukaryotic orthologous groups (KOG) and Homologene for eukaryotes (2–4). Similar initiatives by other groups include the UniProt HAMAP (High-quality Automated and Manual Annotation of microbial Proteomes), KEGG (Kyoto Encyclopedia of Genes and Genomes) orthology groups, The Institute for Genomic Research's (TIGR—now JCVI) TIGRFAMs, ACLAME (A CLAssification of genetic Mobile Elements) protein clusters for mobile elements and MBGD (Microbial Genome Database) clusters (5–9). Related protein sequences need not be analyzed as families of complete sequences and structural studies have led to the formation of numerous domain (generally defined as independently folded 3D structures) databases including the NCBI Conserved Domain Database (CDD), the European Bioinformatics Institute's InterPro database, Pfam (protein families) and SMART (Simple Modular Architecture Research Tool; 10–13) groups. Other databases have been built for analyzing metabolic pathways and systems such as in KEGG or BioCyc to provide more holistic analyses of molecular functions and their interrelationships (14,15). Some databases attempt to find only orthologous clusters; while, others contain clusters of both orthologs and paralogs. Some only contain curated data sets; whereas, others contain both curated and uncurated families. Some are frequently updated while others are not. Many are routinely used in the analyses of large-scale genomic and metagenomic data sets.

Regardless of the various databases and tools that have been generated over the years to analyze protein function, one major difficulty has been in keeping the annotated data from the growing genomic data sets up-to-date.

*To whom correspondence should be addressed. Tel: +1 301 496 4859; Fax: +1 301 402 9651; Email: klimke@ncbi.nlm.nih.gov

The NCBI Reference Sequence (RefSeq) project that contains nonredundant sets of curated transcripts, gene and protein information in eukaryotic organisms, and gene and protein information in prokaryotes, has been a very successful way to maintain and update annotated data (16). It was realized that annotating protein families as a group was a convenient and efficient way to functionally annotate the increasing numbers of prokaryotic genomes that were being deposited at an increasing rate. The Protein Clusters database (ProtClustDB) has been constructed with two goals in mind: first, to routinely update RefSeq genomes with curated gene and protein information from ProtClustDB 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 (4). In addition, one of the most important sources of information, the scientific literature itself, where important experimentally verified functions are reported, is routinely parsed for existing or potential connections to genes/proteins and connected to each cluster. The NCBI ProtClustDB is available at http://www.ncbi.nlm.nih.gov/sites/entrez?db=proteinclusters.

DATA CONTENT AND ORGANIZATION

Clustering

The ProtClustDB consists of proteins encoded by complete chromosomes and plasmids from the RefSeq collection in four sets or groups: prokaryotes, phages, chloroplasts and mitochondria. Clusters from each set are created separately and given different accession prefixes (Table 1). Proteins are compared by sequence similarity using BLAST all against all (*E*-value cutoff 10E-05; effective length of the search space 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 the cluster will have to protein A. There are no cutoffs used during the clustering procedure, or strict requirements for clusters of orthologous groups, or any check on phylogenetic distance. The initial set of uncurated clusters created in 2005 has been used as a starting point for curation and has been updated quarterly since that time. 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 the smaller clusters. The first web release of ProtClustDB was in 2007 (4).

Related clusters

Starting with existing curated and uncurated cluster membership as a basis, related clusters are calculated using all-against-all BLASTP and RPS-BLAST of proteins against profiles in the CDD. Unlike initial cluster creation, calculation of cluster relationships includes clusters from all four taxonomic groups. Currently, related clusters consists of sets of disjoint clusters where for every pair of clusters A and B in a set, every protein in cluster A is related to every protein in cluster B. For a pair of proteins to be related, there must be a BLASTP alignment between the two (*E*-value cutoff 0.001) that covers at least 80% of the length of the shorter of the two proteins, and either both proteins do not have any RPS-BLAST matches (*E*-value cutoff 0.01) or at least 80% of the domains are shared between two proteins with the center point of the domain alignment on the domain for the two proteins within 25 residues of each other. Sets are made from pairs of clusters satisfying above conditions in a greedy fashion based on the product of the number of proteins in the two pairs or sets made so far.

Cluster curation

Curated information includes functional annotation for genes and proteins, Enzyme Commission numbers

Table 1. Statistics for Entrez Protein Clusters—May, 2008

Cluster type[a]	Accession prefix[b]	Nucleotide sequences[c]	Proteins[d]	Proteins in clusters[e]	Clusters[f]	Curated clusters[g]	Proteins in curated clusters[h]	Publications[i]
Prokaryotic	PRK/CLS	2024	2 248 112	1 708 872	281 861	6524	356 618	2611/55 894
Chloroplast	CHL/CLSC	131	12 529	10 706	646	163	9106	29/2508
Mitochondrial	MTH/CLSM	1213	15 763	10 118	595	161	8968	1/94
Phage	PHA/CLSP	438	28 355	14 508	2284	332	1821	299/6518
Totals	N/A	3806	2 304 759	1 744 204	285 386	7180	376 513	2939/65 016

[a]Major groups (clustered separately).
[b]Prefix (curated/uncurated) for each group.
[c]Total number of nucleotide sequences from RefSeq genomes/plasmids.
[d]Total number of proteins encoded by all nucleotide records in #3.
[e]Total number of proteins within clusters.
[f]Total number of clusters.
[g]Total number of curated clusters.
[h]Proteins contained within the curated cluster set.
[i]Number of unique publications (curated/all types) across all clusters. The nonredundant total includes publications describing functions in multiple cluster groups.

identifying enzymatic function, publications describing function, and protein content (17). Protein names are the only curation that is required as they are the most prominent annotation-describing function. Gene names are often added but are not required. Functional descriptions provide more detailed information, and curators typically add publications that describe experimental evidence of function for at least one member in each cluster or in a related cluster. Clusters may also be joined together into larger curated clusters, or clusters may be split if there is evidence of functional divergence between protein members, or if significant numbers of paralogous members exist within a cluster. Curators also use multiple alignments to alter start sites for coding sequences in the RefSeq collection in order to correct miscalled translation initiation sites which typically had resulted in truncation of a shared conserved domain. Cluster curation not only provides updated functional information in ProtClustDB, but it is also used to transfer annotation to all protein members, and this is reflected in the genomic RefSeq records in Entrez Nucleotide, Genome, Protein and Gene. Curated domains are mirrored in the CDD. The curated and some chosen uncurated clusters are also used as sources of protein names for function annotation for submissions to the prokaryotic genome automatic annotation pipeline (PGAAP, http://www.ncbi.nlm.nih.gov/genomes/static/Pipeline.html).

The number of clusters is quite large as shown in Supplementary Figure S1. Many clusters are composed of only a small number of proteins, especially in the uncurated set, which reflects the conservative clustering requirements. The majority of proteins in uncurated clusters are in clusters with 10 or less protein members, while in the curated sets most proteins are in clusters with more than 10 proteins (data not shown). During curation many smaller clusters may be joined together to form one large curated cluster, which is why the curated set trends towards larger cluster sizes. This reflects the trend of more conserved and larger protein families tending to be well studied with experimentally determined functions than smaller protein families. Cluster size also reflects both the bias in sequencing efforts in certain taxonomic branches and the bias of certain model organisms in those branches. For example, curation centered around clusters that contain *Escherichia coli* K-12 proteins includes information from curated databases such as EcoGene and EcoCyc (18,19). This means that there are many curated clusters in the Enterobacteriaceae taxonomic branch as compared to other organism groups and spikes in the graph (Supplementary Figure S1A3) around cluster sizes of 40 members represent many curated clusters from this branch (data not shown). Constant curation has led to an increase in the number of proteins contained within curated clusters for each release.

One common source of clustered protein information that is utilized in microbial genomic studies is the COG database. A comparison of COG associations with protein clusters (protein members with COG association, either through the 66 genomic constituents for the last COG release, or via the COGs mirrored in CDD) shows that many cluster functional categories are associated with uncurated clusters but that there are curated clusters for almost all functional categories including previously uncharacterized COGs (Supplementary Table S1; Figure S2). There are a few exceptions in some very small COG sets, or in COGs that are not present in prokaryotes. Curation of what were previously 'function unknown' COG categories has led to improvement of protein cluster and RefSeq annotation, for example, the identification of COG1892 as a phosphoenolpyruvate carboxylase has now been used to curate a cluster [PRK13655; (20)]. Information on protein function as reported in the literature will help to improve curated clusters and functional annotation on the RefSeq genomes, allowing all scientists to obtain the most up-to-date information on molecular functions.

Internal and external links

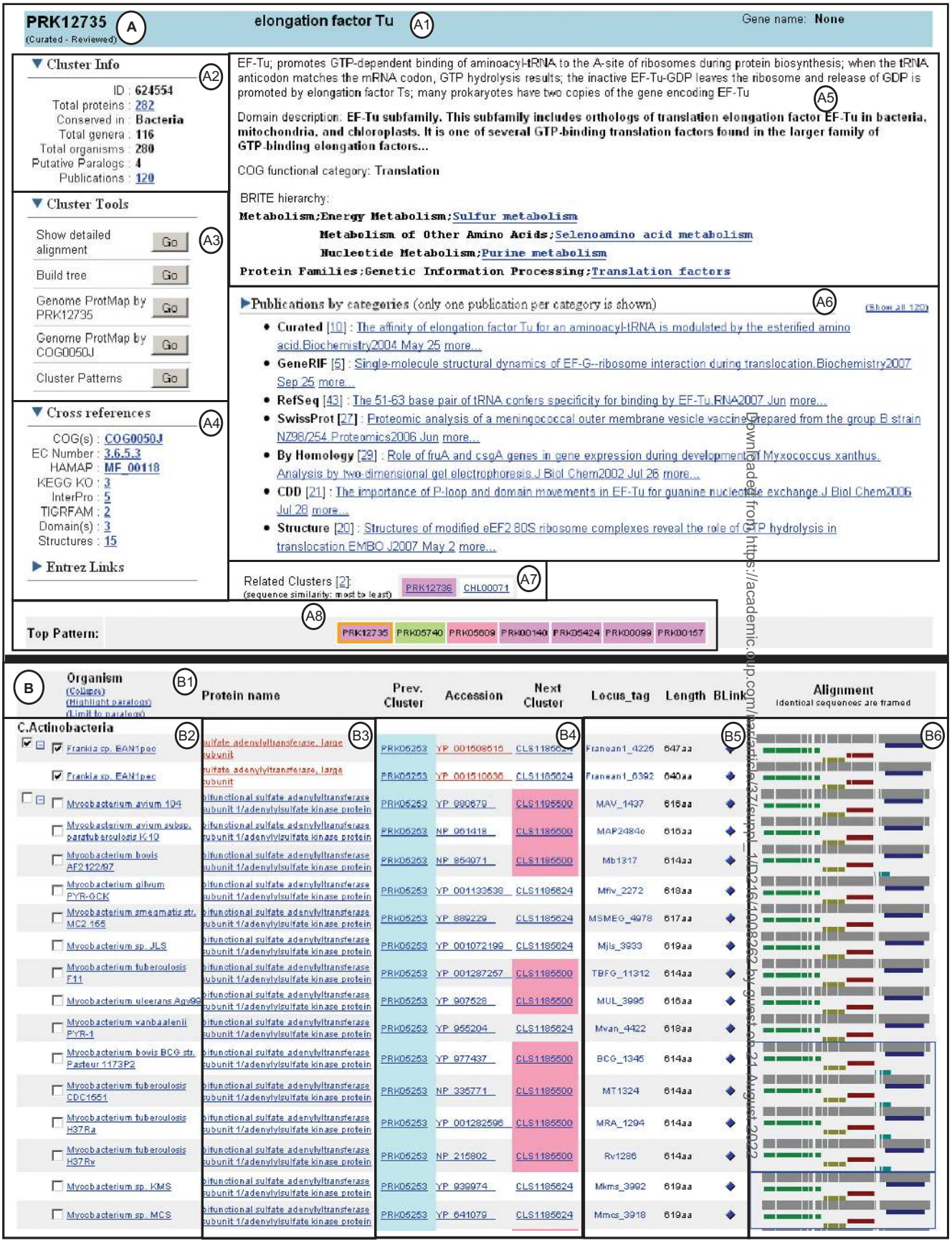
Along with the curated information added above, a vast amount of information is collected from NCBI and external resources and added as cross references. All proteins and genes comprising each cluster, and the associated links to them, are collected for each cluster and the links include: clusters of orthologous groups (COG), domain and protein families (CDD), and structures from the Protein Data Bank (PDB) via Molecular Modeling Database (MMDB) mirror (21). PubMed identifiers are collected from all cluster members from many resources and from related proteins not in RefSeq identified via sequence similarity searches of the nonredundant protein BLAST databases. Other cross-references are made to ACLAME, EC Numbers, HAMAP, KEGG orthology groups and BRITE hierarchy information, InterPro and TIGRFAMs.

DATA ACCESS

Entrez protein clusters

The first public release of the ProtClustDB via NCBI's Entrez interface was in April, 2007 and initially consisted of only prokaryotic clusters (4). Since that time, organellar and bacteriophage clusters have been added. Quarterly updates have taken place with the addition of new proteins and newly created clusters to the existing data set. The statistics representing the latest release (May 2008) are shown in Table 1. The largest set of clusters includes the prokaryotic chromosomes and plasmids, followed distantly by the phage, mitochondrial and chloroplast groups.

The ProtClustDB is available in NCBI's Entrez system (http://www.ncbi.nlm.nih.gov/sites/entrez?db=proteinclusters), the help document provides more detail (http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=helpcluster.chapter.helpcluster) and the data are also available via File Transfer Protocol (FTP) (see subsequently). The Entrez system provides a mechanism for the search, retrieval and linkage between protein clusters and other NCBI databases as well as external resources (4). Clusters can be searched by general text terms, and also by specific protein or gene names. Clusters can also be filtered by curated or noncurated sets, by taxonomic



groupings, and by more powerful searches such as average protein length, or cluster size, and other combinations of queries are possible.

A typical cluster overview is shown in Figure 1. The display is partitioned into two halves, cluster information and a table of proteins. The top half contains curated and automatically collected information for the complete cluster and content analysis tools. The bottom half displays each protein in a row along with a schematic of an automatically made and noncurated multiple alignment created using the multiple sequence alignment program Muscle [using fastest parameters except for -maxiters 99; (22)]. A number of analytical tools are available from the display page that provide a more detailed analysis of the multiple alignment, phylogenetic tree and the genome neighborhood as shown in Figure 2.

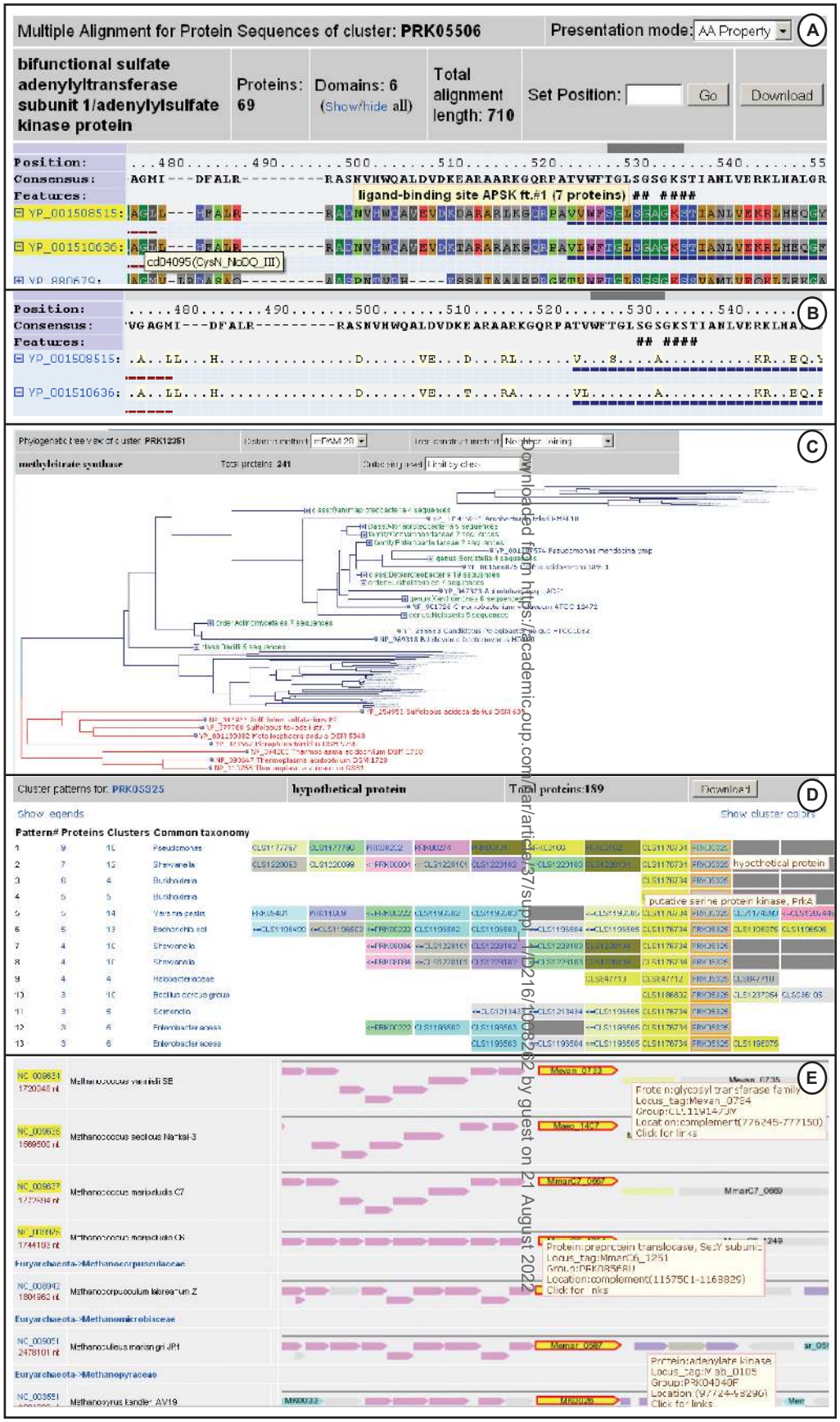
Unlike other clustered protein databases, the ProtClustDB provides access to both curated and uncurated clusters. The curated clusters provide uniform names and annotation that have been curated by NCBI curators. There are three status levels indicating the amount of curation applied: provisional for minimal curation, validated for medium level and reviewed for extremely confident curation. Status levels change during cluster curation.

One of the most important aspects of ongoing annotation efforts is the experimental derivation of function that is reported in the literature. Only a small number of publications are connected directly to protein and nucleotide sequences in the public databases. In order to help spur that effort, many publication links are automatically collected from a variety of NCBI resources: Entrez Gene GeneRIFs (gene reference into function) as well as other types, publications added by external collaborators or by NCBI curators, from RefSeq and related proteins submitted to the primary data archives (INSD—GenBank/DDBJ/EMBL), curated publications added to UniProt records, publications linked to structure records and from CDD. Curated publications are also added directly to some clusters. All publications are prominently displayed in category groups and allow rapid exploration of the literature information space directly through the publications associated with a cluster and indirectly via pre-calculated related articles and cited articles in PubMed and PubMed Central (4). As shown in Table 1, the number of curated publications (2939) and the total number of publication to cluster links (65 016) is a large, though obviously not comprehensive, set of published information. This important ongoing effort is expected to increase the number of publication links in the future.

Cluster tools

Cluster tools are available to analyze and explore clusters through the examination of the multiple alignment, phylogenetic tree and genomic neighborhoods (Figure 2). The multiple alignment displays specific regions covered by conserved domains and features derived from CDD (imported from structural information) and the ability to download the entire alignment. The phylogenetic tree display utilizes many of the advances used for display of large data sets for the NCBI Flu database such as compaction without loss of visual information (23). The ProtMap and cluster patterns show genomic neighborhoods either for all genomes (ProtMap) or in a taxonomically collapsed view by conserved patterns of clusters. ProtMap can also be used to show the genomic neighborhood by COG or by VOG [viral COGs; (4)]. All cluster tools enable detailed exploration of a single cluster in a greater context, and is especially useful for uncurated clusters or clusters where no function has yet been described for the cluster members.

Figure 1. Cluster overview display. (**A**) Overview of one of the curated elongation factor Tu clusters (PRK12735). All expandable panels are marked with an arrowhead. (A1) Cluster Accession, curation status and protein name, either curated or automatically chosen from existing names for uncurated clusters. Curated gene names would appear at the right. (A2) The cluster info panel includes basic statistics for the cluster including protein, paralog, genera and publication counts. (A3) Cluster tool panel for launching separate analysis tools (shown in detail in Figure 2). (A4) Cross-references to NCBI and external databases from both curated and automatically collected information. NCBI links include references to the COG, conserved domain (CDD) and structure (MMDB) and other Entrez databases (collapsed in current view—gene, protein, nucleotide, genome, PubMed and taxonomy). External links are described in the text. When there is more than one link in a category, the full list is shown when clicking on that particular category and a single link can be chosen. (A5) Curated functional descriptions, domain description from NCBI CDD, COG functional category and KEGG BRITE hierarchy. (A6) Publication categories. The full set of publications is available as a link to PubMed for the full set or each subset separately. Publications may occur in multiple categories. (A7) Related clusters section shows up to 10 related curated and uncurated clusters from all four cluster groups. The full nonredundant set is available from the link showing the total number of related clusters. (A8) Top cluster pattern. The pattern tool collects patterns of conserved clusters (present in at least three genomes) with the most conserved pattern displayed on the overview page. All patterns are available by clicking on the image and from the cluster tool (Figure 2D). (**B**) Protein table for curated cluster PRK05506 (bifunctional sulfate adenylyltransferase subunit 1/adenylylsulfate kinase protein). The list of proteins is displayed below the cluster overview. (B1) Column headers. This section includes tools to control the list of proteins such as collapsing all organism groups (this can also be done individually for each group). Paralogs (two or more proteins encoded by the same nucleotide sequence) can be highlighted in yellow, or the entire protein table can be limited to paralogs only. (B2) List of organism groups and organisms. Checkboxes are used to highlight groups or individual proteins which can be used to broadcast selections to highlight proteins in the cluster tool displays (Figure 2). Two proteins from *Frankia* genomes have been selected in order to highlight them in the alignment tool (Figure 2A). (B3) The list of current protein names reflects the current set of names from RefSeq proteins. Once all proteins in a cluster are updated with the curated name then all protein names will be the same (as they are in this image). (B4) Protein RefSeq Accession Number and local genomic neighborhood. Genes encoding a protein in the current cluster are examined in both upstream and downstream flanking genes in each genome to check for cluster assignment. Genes in a cluster are shown with that cluster accession, those clusters with a COG association are shown color-coded by functional category. Unclustered genes or RNA genes or pseudogenes are not shown at all. This provides a quick snapshot of the local genomic neighborhood for each gene in the cluster. In this image, all upstream genes encode proteins that belong to curated cluster PRK05253 (sulfate adenylyltransferase subunit 2). (B5) Links to Entrez Gene by locus tag (unique gene identifier), the protein length and Blink results for each protein [BLast link—pre-computed BLAST results for proteins—blue diamond; (24)]. (B6) Alignment schematic. Aligned regions are shown as shaded gray bars with domain information drawn as color-coded bars below each protein (the color is randomly chosen). Sequences that are absolutely identical to each other are framed with a box.



Sequence search

Sequence searches against the protein cluster database are available in two types. The first, Concise Protein BLAST can be used for both protein and nucleotide searches using BLASTP or BLASTX, respectively (http://www.ncbi.nlm.nih.gov/genomes/prokhits.cgi). The concise database contains proteins in all clusters, both curated and uncurated, as well as all nonclustered proteins. From the clustered set, a single random representative at the genus level is chosen in order to reduce the data set. Therefore, results are available more rapidly and the results that are returned provide a broader taxonomic range due to this data reduction.

The second search type utilizes 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 as well as being 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 are available from FTP (see subsequently).

FTP

Quarterly releases are available for download from the FTP directory (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/CLUSTERS/) by date and by the four taxonomic groups. Flatfiles are created for all curated clusters containing the full set of annotated information at the time of release and much of the automatically collected information. Uncurated clusters are in a separate file (gi to cluster ID). Additional files include all publication links, all nucleotide and taxonomic accessions and identifiers, pre-made alignments and the pre-calculated PSSMs for curated clusters. The concise protein BLAST database for prokaryotes is also available (as a subset of the larger nr database).

CONCLUSIONS AND PERSPECTIVES

The NCBI ProtClustDB consists of aggregates of related protein sequences, some of which have been curated with functional annotation, and/or contain automatically collected information. All clusters provide rich sources of information for the analysis of single protein families, or in large-scale analyses across many genomes and quarterly updates will provide up-to-date information for both curated and uncurated sets for the scientific community.

FUTURE DIRECTIONS

Additional subsets of clusters including species- and other taxonomic-level clusters will be used to generate separate displays and made available to represent more or less conserved functions along all taxonomic branches. Supercluster tools will be developed to provide analyses of related clusters. Related clusters are undergoing optimization to find the best criteria to establish sets of relationships. Ongoing updates to the cluster content and the analytical tools will continue to be made at quarterly intervals.

Figure 2. Cluster tools. (**A**) Detailed multiple alignment view for cluster PRK05506 (bifunctional sulfate adenylyltransferase subunit 1/adenylylsulfate kinase protein—Figure 1B). The detailed alignment view provides the capability to display the alignment that is color-coded by conserved amino acid property, which highlights residues at 80% or greater in the following redundant groups: aromatic (FHWY); aliphatic (ILVA); hydrophobic (ACFILMVWY); alcohol-containing (STC); charged (DEHKR); positive (HKR); negative (DE); polar (CDEHKNQRST); tiny (AGS); small (ACDGNPSTV); or bulky (EFIKLMQRWY); or by consensus mode as shown in the next panel. (**B**) The top panel includes information and controls for the alignment as well as a download button (FASTA + gap). Domains and features aligned against each protein (drawn as colored bars under the protein sequence) are from CDD. In this example, two domains are displayed in the alignment drawn as colored boxes below the sequence for the two highlighted proteins from *Frankia*: cd04095, domain II of ATP sulfurylase, brown on the left and cd0207—adenosine 5′-phosphosulfate kinase, blue on the right, with a ligand-binding site in the feature row above the protein sequences. (**C**) Phylogenetic tree for PRK12351 (methylcitrate synthase). At the top is the toolbar with information and controls for distance method, tree construction method and the collapse level (by taxonomic rank). Below is the tree which in this image has been rerooted, showing archaeal proteins highlighted in red (in this case from checkboxes from the protein table for this cluster) and expanded to show every leaf. Transformations of the tree can be done by clicking on the tree itself (reroot, squeeze, collapse and expand). (**D**) Cluster pattern view for PRK05325 (hypothetical protein). The pattern tool allows for exploration of conserved gene neighborhoods. Whereas, the protein table and ProtMap shows the complete genomic region around each gene encoding a protein in a cluster, the pattern tool collects conserved patterns that occur in three or more genomes, in a maximum window of 40 genes upstream or downstream. The most conserved pattern is shown at the top (and on the overview page—Figure 1A8) and the number of conserved proteins which is the number of sequences contributing to the same pattern (which may be from the same nucleotide sequence if present as paralogs in the same cluster), number of clusters in the conserved pattern and common taxonomic node are shown in the table to the left of the patterns. The pattern itself shows all clusters in each pattern and is pseudo-aligned, with the same cluster in each row aligned. Clusters are color-coded according to COG functional categories and the accession is linked to the cluster, the cluster pattern or the ProtMap for that particular cluster. Gray boxes indicate an insertion into the pseudo-alignment for alignment purposes and does not reflect a cluster (gene/protein) at that position in the genome. The size of each box is not proportional to the size of the gene as the size of the arrows is in ProtMap. The gene neighborhood around the genes encoding the hypothetical proteins for PRK05325 (no function yet determined) show conservation of association with putative serine protein kinases (the yellow category apparently involved in signal transduction—a set of uncurated clusters encoded by genes 5′ of the genes encoding proteins in PRK05325). (**E**) Limited ProtMap view for PRK08568 (preprotein translocase subunit SecY). The ProtMap view shows the full gene neighborhood in a limited horizontal window, unlike the cluster pattern tool which shows a more condensed and taxonomically conserved view of the same information but with a potentially wider window. Note that the genes are drawn to scale in this view. In this example, the *Methanococcus* spp. RefSeq Nucleotide Accession Numbers are highlighted in yellow on the left to show that the *secY* gene (cluster PRK08568) is found upstream of a glycosyl transferase encoding gene (CLS1191473—color-coded yellow for cell wall biogenesis); whereas, in most other organisms *secY* is upstream of adenylate kinase (PRK04040—colored blue for nucleotide transport and metabolism). Note that PRK04040 contains a large set of contributing sequences that are not shown in the image for brevity. The pattern tool can be used to control the display of the ProtMap, directing the display to only show the ProtMap for a particular pattern.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Funding for open access charge: Intramural Research Program of the National Institutes of Health; National Library of Medicine.

Conflict of interest statement. None declared.

REFERENCES

1. Fleischmann,R.D., Adams,M.D., White,O., Clayton,R.A., Kirkness,E.F., Kerlavage,A.R., Bult,C.J., Tomb,J.F., Dougherty,B.A., Merrick,J.M. *et al*. (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, **269**, 496–512.
2. Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.
3. Tatusov,R.L., Fedorova,N.D., Jackson,J.D., Jacobs,A.R., Kiryutin,B., Koonin,E.V., Krylov,D.M., Mazumder,R., Mekhedov,S.L., Nikolskaya,A.N. *et al*. (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.
4. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S. *et al*. (2007) Database resources of the national center for biotechnology information. *Nucleic Acids Res*., **35**, D5–D12.
5. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al*. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*., **31**, 365–370.
6. Leplae,R., Hebrant,A., Wodak,S.J. and Toussaint,A. (2004) ACLAME: a CLAssification of mobile genetic elements. *Nucleic Acids Res*., **32**, D45–D49.
7. Okuda,S., Yamada,T., Hamajima,M., Itoh,M., Katayama,T., Bork,P., Goto,S. and Kanehisa,M. (2008) KEGG Atlas mapping for global analysis of metabolic pathways. *Nucleic Acids Res*., **36**, W423–W426.
8. Selengut,J.D., Haft,D.H., Davidsen,T., Ganapathy,A., Gwinn-Giglio,M., Nelson,W.C., Richter,A.R. and White,O. (2007) TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes. *Nucleic Acids Res*., **35**, D260–D264.
9. Uchiyama,I. (2007) MBGD: a platform for microbial comparative genomics based on the automated construction of orthologous groups. *Nucleic Acids Res*., **35**, D343–D346.
10. Marchler-Bauer,A., Anderson,J.B., Derbyshire,M.K., DeWeese-Scott,C., Gonzales,N.R., Gwadz,M., Hao,L., He,S., Hurwitz,D.I., Jackson,J.D. *et al*. (2007) CDD: a conserved domain database for interactive domain family analysis. *Nucleic Acids Res*., **35**, D237–D240.
11. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bork,P., Buillard,V., Cerutti,L., Copley,R. *et al*. (2007) New developments in the InterPro database. *Nucleic Acids Res*., **35**, D224–D228.
12. Finn,R.D., Tate,J., Mistry,J., Coggill,P.C., Sammut,S.J., Hotz,H.R., Ceric,G., Forslund,K., Eddy,S.R., Sonnhammer,E.L. *et al*. (2008) The Pfam protein families database. *Nucleic Acids Res*., **36**, D281–D288.
13. Letunic,I., Copley,R.R., Schmidt,S., Ciccarelli,F.D., Doerks,T., Schultz,J., Ponting,C.P. and Bork,P. (2004) SMART 4.0: towards genomic data integration. *Nucleic Acids Res*., **32**, D142–D144.
14. Kanehisa,M., Araki,M., Goto,S., Hattori,M., Hirakawa,M., Itoh,M., Katayama,T., Kawashima,S., Okuda,S., Tokimatsu,T. *et al*. (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res*., **36**, D480–D484.
15. Caspi,R., Foerster,H., Fulcher,C.A., Kaipa,P., Krummenacker,M., Latendresse,M., Paley,S., Rhee,S.Y., Shearer,A.G., Tissier,C. *et al*. (2008) The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res*., **36**, D623–D631.
16. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*., **35**, D61–D65.
17. Bairoch,A. (2000) The ENZYME database in 2000. *Nucleic Acids Res*., **28**, 304–305.
18. Rudd,K.E. (2000) EcoGene: a genome sequence database for Escherichia coli K-12. *Nucleic Acids Res*., **28**, 60–64.
19. Karp,P.D., Keseler,I.M., Shearer,A., Latendresse,M., Krummenacker,M., Paley,S.M., Paulsen,I., Collado-Vides,J., Gama-Castro,S., Peralta-Gil,M. *et al*. (2007) Multidimensional annotation of the Escherichia coli K-12 genome. *Nucleic Acids Res*., **35**, 7577–7590.
20. Ettema,T.J., Makarova,K.S., Jellema,G.L., Gierman,H.J., Koonin,E.V., Huynen,M.A., de Vos,W.M. and van der Oost,J. (2004) Identification and functional verification of archaeal-type phosphoenolpyruvate carboxylase, a missing link in archaeal central carbohydrate metabolism. *J. Bacteriol*., **186**, 7754–7762.
21. Wang,Y., Addess,K.J., Chen,J., Geer,L.Y., He,J., He,S., Lu,S., Madej,T., Marchler-Bauer,A., Thiessen,P.A. *et al*. (2007) MMDB: annotating protein sequences with Entrez's 3D-structure database. *Nucleic Acids Res*., **35**, D298–D300.
22. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*., **32**, 1792–1797.
23. Bao,Y., Bolotov,P., Dernovoy,D., Kiryutin,B., Zaslavsky,L., Tatusova,T., Ostell,J. and Lipman,D. (2008) The influenza virus resource at the national center for biotechnology information. *J. Virol*., **82**, 596–601.
24. Wheeler,D.L., Church,D.M., Lash,A.E., Leipe,D.D., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Tatusova,T.A., Wagner,L. *et al*. (2001) Database resources of the national center for biotechnology information. *Nucleic Acids Res*., **29**, 11–16.