Identifying Protein Domains with the Pfam Database

Penny Coggill,[1] Robert D. Finn,[1] and Alex Bateman[1]

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom

ABSTRACT

Pfam is a database of protein domain families, with each family represented by multiple sequence alignments and profile hidden Markov models (HMMs). In addition, each family has associated annotation, literature references, and links to other databases. The entries in Pfam are available via the World Wide Web and in flatfile format. This unit contains detailed information on how to access and utilize the information present in the Pfam database, namely the families, multiple alignments, and annotation. Details on running Pfam, both remotely and locally are presented. *Curr. Protoc. Bioinform.* 23:2.5.1-2.5.17. © 2008 by John Wiley & Sons, Inc.

Keywords: protein domain • HMM • protein family • superfamily • sequence alignment • sequence analysis

INTRODUCTION

Millions of protein sequences are now known and the deluge of data shows no sign of slowing. The sequence analysis of proteins may seem like a perpetual task, but the majority of proteins do appear to fall into a few thousand protein families (Chothia, 1992). Very often, these families are representative of proteins at the domain level, where domains are discrete structural-functional units that are found in many different proteins in many different protein contexts. Many of these domains are found to be distantly related and often have similar functions. Pfam is a database of such protein domain families (Sonnhammer et al., 1997*;* Finn et al*.,* 2008*)*, with each family represented by a multiple sequence alignment and a profile hidden Markov model (HMMs; see *APPENDIX 3A*). In addition, each family has associated annotation, literature references, and links to other databases. Pfam also contains sets of related families called clans. The entries in Pfam are available via the Web and in flatfile format.

The use of Pfam by molecular biologists as a protein information resource and analysis tool is widespread. Genome and metagenomic sequencing projects, including the Global Ocean Survey (Yooseph et al., 2007), have used Pfam extensively for large-scale functional annotation of genomic data, while smaller groups devoted to studying a single protein or biochemical pathway have frequently used Pfam for their analyses. The multiple sequence alignments around which Pfam families are built are important for understanding both protein structure and protein function, and form the basis for techniques such as secondary structure-prediction, fold-recognition, and phylogenetic analysis, and can aid in mutation design.

This unit contains detailed information on how to access and utilize the information present in the Pfam database. Details on running Pfam both remotely via the Web (Basic Protocol) and locally (Alternate Protocol 1) are presented; for using the latter the *APPENDIX* at the end of the unit gives a description of the files available through the Web site. This unit also includes a brief discussion on analyzing genomic DNA with Pfam (Alternate Protocol 2).

Current Protocols in Bioinformatics 2.5.1-2.5.17, September 2008
Published online September 2008 in Wiley Interscience (www.interscience.wiley.com).
DOI: 10.1002/0471250953.bi0205s23

BASIC PROTOCOL

ANALYZING A PROTEIN SEQUENCE WITH Pfam VIA THE WEB

The most common and primary use of the Pfam database is to determine what domains are present in a protein of interest. This section describes several approaches for carrying out such an analysis. In a typical session any one of them, or several in combination, may be used, depending on the kinds of questions asked and the types of information sought. Pfam contains not only protein domains, but also other regions of similarity, and within this unit the terms family and domain are used equivalently.

Necessary Resources

Hardware

Workstation with network connection

Software

Javascript-enabled browser (e.g., Mozilla, Firefox, or Internet Explorer)

Files

Protein sequence of interest in appropriate format (e.g., FASTA; *APPENDIX 1B*)

Access Pfam via the Web

1. Access the Pfam Web site, which is currently available at three locations around the world:

   http://pfam.sanger.ac.uk (U.K.)
   http://pfam.sbc.su.se (Sweden)
   http://pfam.janelia.org (U.S.).

   The latest documentation on the content and use of Pfam is available via the Web, and should be referred to alongside this unit. The uses of core functions available at all sites are described in the following sections.

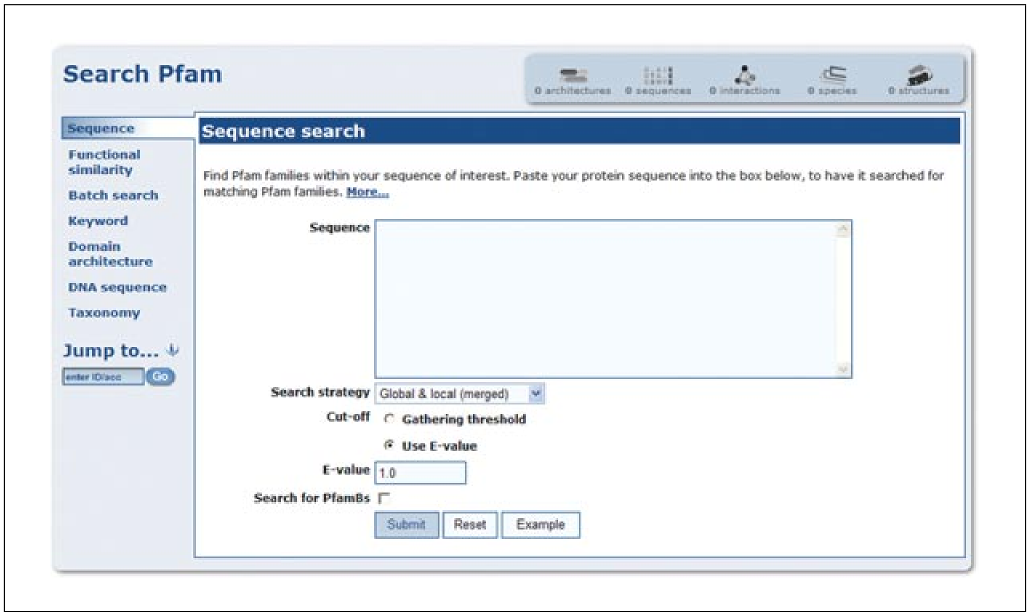
Search strategies

2a. *Search Pfam with a novel protein sequence:* From the Pfam homepage of any of the three mirror-sites, click on either SEARCH on the top menu bar or SEQUENCE SEARCH in the central list. Clicking the former reaches the full Search page (Fig. 2.5.1), whereas clicking the latter reaches a simpler option that can be expanded to the more detailed one. Paste the sequence in the appropriate format into the text box to perform a sequence search of Pfam.

These two pages allow the user to paste the sequence in the appropriate format into the text box, which can be FASTA (see APPENDIX 1B) or bare sequence with no header line. The sequence is searched against the Pfam collection of HMMs.

For most purposes the default settings on the Search page will be sufficient (see Critical Parameters and Troubleshooting). The user can optionally include a search of Pfam-Bs, as well as the default search against Pfam-As (see Background Information). The searches can take from between 5 sec and 2 min to run depending on server load and sequence length, but the user is given an estimated running time, and a progress bar graphic indicates progress.

The results page that appears gives a graphical representation of the domain(s) found on the sequence and a table of the domain matches. Optionally, the user can display pairwise sequence alignments of the query sequence to the relevant HMM. Figure 2.5.2 gives an example of a search output. Clicking on any domain graphic or on the domain name brings up the relevant Family page. The domain graphic will appear just as a gray bar when there is no match (see also Cut-off Thresholds paragraph in the Guidelines for Interpreting Results section).



Search Pfam

Sequence search

Find Pfam families within your sequence of interest. Paste your protein sequence into the box below, to have it searched for matching Pfam families. More...

Sequence

Search strategy Global & local (merged)

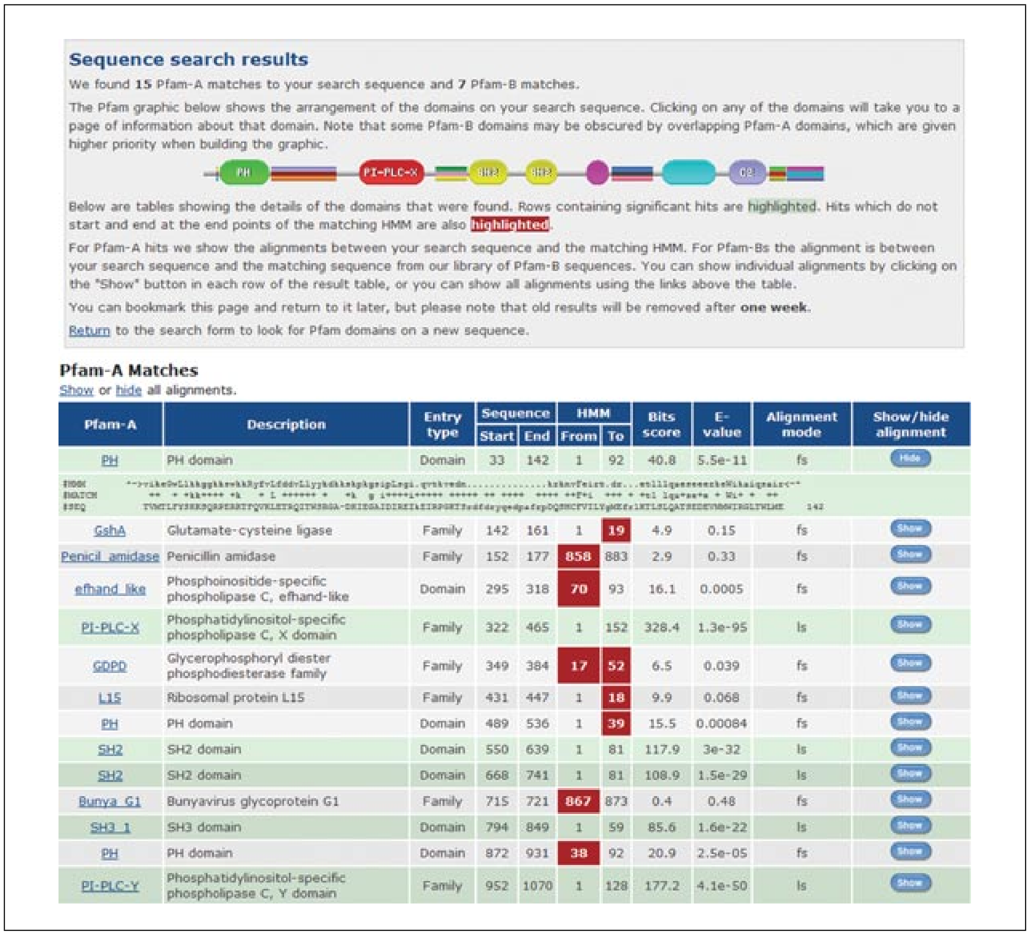
Cut-off ○ Gathering threshold

● Use E-value

E-value 1.0

Search for PfamBs □

Figure 2.5.1 Screenshot of the protein search submission page. The use of this search page is described in detail in the main text.



Sequence search results

We found **15** Pfam-A matches to your search sequence and **7** Pfam-B matches.

The Pfam graphic below shows the arrangement of the domains on your search sequence. Clicking on any of the domains will take you to a page of information about that domain. Note that some Pfam-B domains may be obscured by overlapping Pfam-A domains, which are given higher priority when building the graphic.

Below are tables showing the details of the domains that were found. Rows containing significant hits are highlighted. Hits which do not start and end at the end points of the matching HMM are also highlighted.

For Pfam-A hits we show the alignments between your search sequence and the matching HMM. For Pfam-Bs the alignment is between your search sequence and the matching sequence from our library of Pfam-B sequences. You can show individual alignments by clicking on the "Show" button in each row of the result table, or you can show all alignments using the links above the table.

You can bookmark this page and return to it later, but please note that old results will be removed after **one week**.

Return to the search form to look for Pfam domains on a new sequence.

Pfam-A Matches

Show or hide all alignments.

Pfam-A	Description	Entry type	Sequence		HMM		Bits score	E-value	Alignment mode	Show/hide alignment
			Start	End	From	To				
PH	PH domain	Domain	33	142	1	92	40.8	5.5e-11	fs	Hide
#HMM [illegible] #MATCH [illegible] #SEQ [illegible] 142										
GshA	Glutamate-cysteine ligase	Family	142	161	1	19	4.9	0.15	fs	Show
Penicil_amidase	Penicillin amidase	Family	152	177	858	883	2.9	0.33	fs	Show
efhand_like	Phosphoinositide-specific phospholipase C, efhand-like	Domain	295	318	70	93	16.1	0.0005	fs	Show
PI-PLC-X	Phosphatidylinositol-specific phospholipase C, X domain	Family	322	465	1	152	328.4	1.3e-95	ls	Show
GDPD	Glycerophosphoryl diester phosphodiesterase family	Family	349	384	17	52	6.5	0.039	fs	Show
L15	Ribosomal protein L15	Family	431	447	1	18	9.9	0.068	fs	Show
PH	PH domain	Domain	489	536	1	39	15.5	0.00084	fs	Show
SH2	SH2 domain	Domain	550	639	1	81	117.9	3e-32	ls	Show
SH2	SH2 domain	Domain	668	741	1	81	108.9	1.5e-29	ls	Show
Bunya_G1	Bunyavirus glycoprotein G1	Family	715	721	867	873	0.4	0.48	fs	Show
SH3_1	SH3 domain	Domain	794	849	1	59	85.6	1.6e-22	ls	Show
PH	PH domain	Domain	872	931	38	92	20.9	2.5e-05	fs	Show
PI-PLC-Y	Phosphatidylinositol-specific phospholipase C, Y domain	Family	952	1070	1	128	177.2	4.1e-50	ls	Show

Figure 2.5.2 Example search outputs from the server for a query sequence of ~1300 residues. The site was queried with the UniProt identifier VAV_HUMAN. In the example, graphical representation of the domains with the query sequence is shown with, below, a table giving the domain positions and match scores.

For searching a large number of proteins against Pfam it may be preferable to search locally (see Alternate Protocol 1). Otherwise, the Web site offers the facility for uploading text files of up to 1000 protein sequences in FASTA format via the Batch Search link on the left-hand menu bar on the SEARCH page. The sequences are then searched against the Pfam HMMs and results are e-mailed back to the user.

2b. *Search Pfam with a UniProt or Genbank identifier or number*: From the homepage click on the central VIEW A SEQUENCE link. Type into the small text input box the UniProt or Genbank accession number or identifier, and click on the Go button.

Entering the protein accession number (e.g., `P15498`, `1322900`*) or identifier (e.g.,* `VAV_HUMAN`, `CAA96955.1`*) is the fastest way to search Pfam, as the authors have precalculated results for profile HMM matches over UniProt and Genbank.*

This will open the Protein page for that accession, which resembles the Family page in format but has some unique, protein-related tabs.

One of the key tabs is the Features link on the left menu bar. Clicking this tab brings up graphical representation of all the additionally annotated features of the domain(s) on the protein, such as disulfide bridges, phosphorylation sites, active sites (Mistry et al., 2007), etc, if these are known. Features from other databases can also be switched on, if available.

2c. *Search Pfam with a PDB protein structure identifier:* From the homepage enter a PDB structure identifier by clicking on the central VIEW A STRUCTURE link and pasting the identifier (e.g., `2bcd`) into the text box.

This search will open a Structure page that, again, resembles the Family page in format but has both cross-links to other structure databases and some unique structure-related tabs. The Domain organization tab shows the domains on the sequence for that structure. The Sequence mapping tab shows sequences and Pfam domains that correspond to this structure. Clicking the View structure tab brings up options to view the structure with Jmol, AstexViewer, or SPICE.

2d. *Search Pfam for protein families related to some word or phrase:* Go to the homepage and type a keyword into the search box in the top right-corner of the page or click on the central KEYWORD SEARCH link. Enter in the word or words to be searched, noting that whitespace will automatically be replaced by the logical operator `AND`, such that the term `RNA polymerase` becomes `RNA AND polymerase`.

This search allows users to query the text contained in all sections of Pfam. The results are ordered according to the number of information types that the query matches to, as listed at the top of the results page.

2e. *Search Pfam with the names or identifiers of known Families or Clans:* The database can be searched with both names and identifiers of Families or Clans from the homepage by clicking on the central VIEW A PFAM FAMILY or VIEW A CLAN and entering the names in the text boxes.

This facility allows rapid access to the Protein or Clan pages (Clan pages are of similar format and functionality to the Family pages). Entries can be in any of the following formats: CBS or PF00571 for Pfam-A entries; PB137753 for known Pfam-B entries from the most recent release, and Kazal or CL0005 for a known clan.

2f. *Search Pfam for families within different Proteomes:* From the homepage, click on the BROWSE field on the top menu bar, and a page with alphabetic listings of all Families, Clans, and Proteomes will appear. Selecting any letter will bring up a full list of all the proteomes with sequences in Pfam starting with that letter.

Clicking on any one species brings up a Proteome page (again resembling a Family page in format) that gives the taxonomic lineage, the number of sequences and domains, and coverage; the left-hand tabs display the domain architectures, domain compositions, and structures found in that proteome.

On all Family-type formatted pages there is a handy ***Jump-to*** *box. Pasting any data-type as used for searches 2b, 2c, and 2e above, i.e., except keyword, sequence, or proteome, brings up the appropriate Family, Protein, or Structure page.*

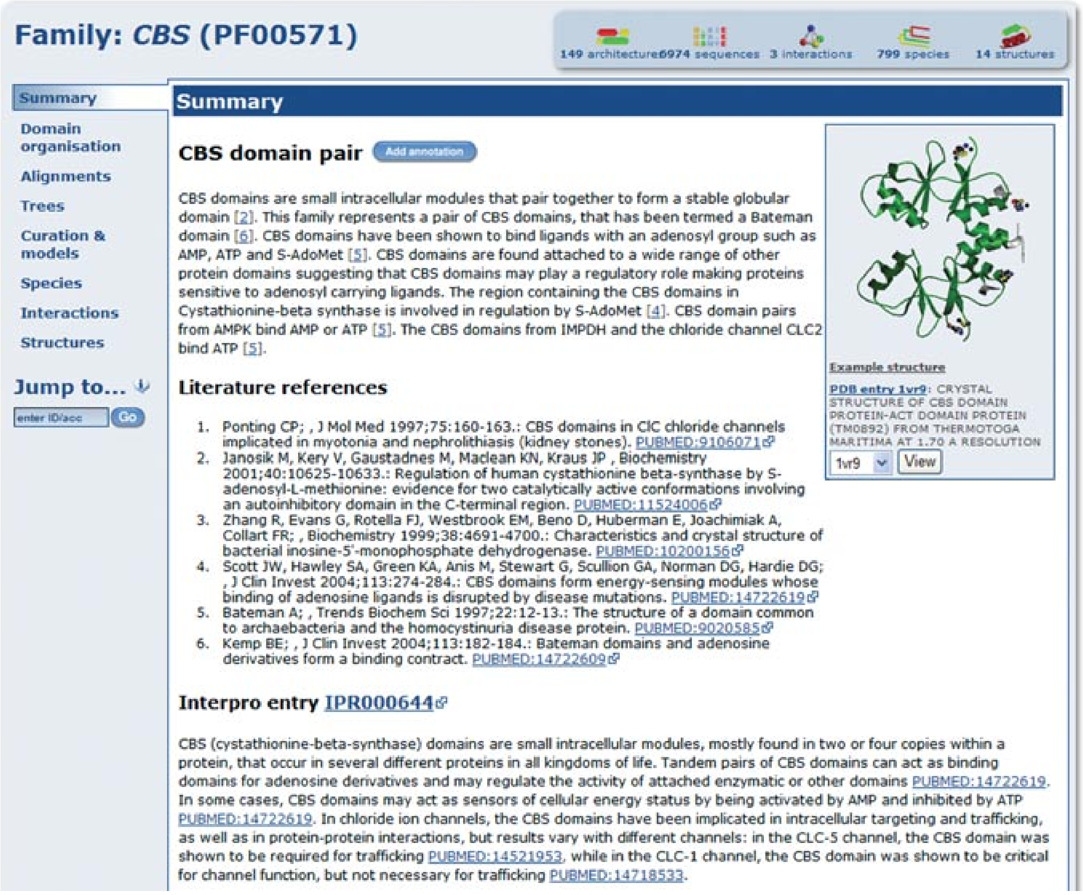
View strategies

View Pfam annotation for a family

3. Click on the Family name or on the graphical depiction of a domain from any Pfam Web page to access the Family page, which allows viewing of the Pfam-family annotation (Fig. 2.5.3).

   The Family pages are the central viewing points for most types of Pfam data on a Pfam family. The front (Summary) page carries the Pfam annotation, PUBMED Literature references, the Interpro annotation, clan details if appropriate, Gene Ontology data, and External database links (see Family annotation for the CBS domain in Fig.2.5.3). Additional information from the flatfile entry is included in the Curation & models tab.

The following sections are available from the left-hand menu bar and each will be expanded further in steps 4, 5, 6, 7, and 8 below, in regards to Alignments, Trees, Curation & Models, Species, and Structures. There is also an Interactions tab that brings up links to all other Pfam families with which this family interacts.



Family: *CBS* (PF00571)

Summary

CBS domain pair

CBS domains are small intracellular modules that pair together to form a stable globular domain [2]. This family represents a pair of CBS domains, that has been termed a Bateman domain [6]. CBS domains have been shown to bind ligands with an adenosyl group such as AMP, ATP and S-AdoMet [5]. CBS domains are found attached to a wide range of other protein domains suggesting that CBS domains may play a regulatory role making proteins sensitive to adenosyl carrying ligands. The region containing the CBS domains in Cystathionine-beta synthase is involved in regulation by S-AdoMet [4]. CBS domain pairs from AMPK bind AMP or ATP [5]. The CBS domains from IMPDH and the chloride channel CLC2 bind ATP [5].

Example structure

PDB entry 1vr9: CRYSTAL STRUCTURE OF CBS DOMAIN PROTEIN-ACT DOMAIN PROTEIN (TM0892) FROM THERMOTOGA MARITIMA AT 1.70 A RESOLUTION

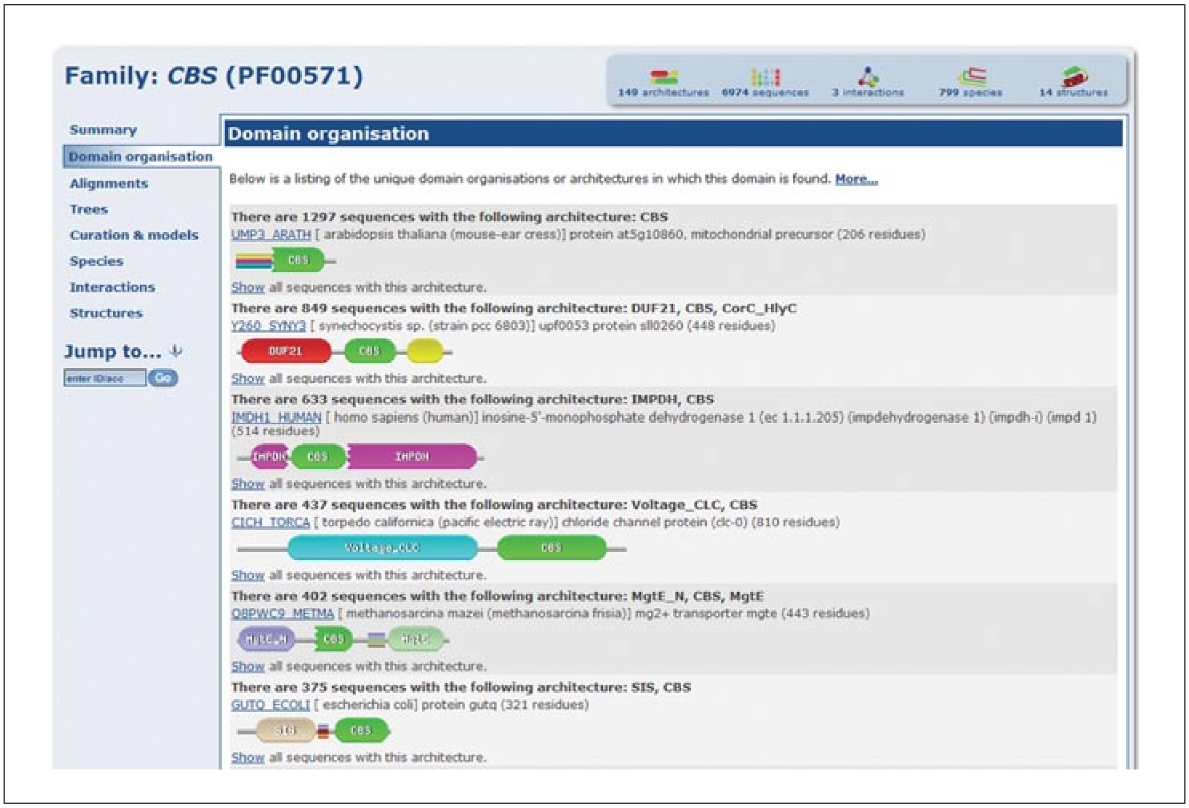
Literature references

1. Ponting CP; , J Mol Med 1997;75:160-163.: CBS domains in CIC chloride channels implicated in myotonia and nephrolithiasis (kidney stones). PUBMED:9106071
2. Janosik M, Kery V, Gaustadnes M, Maclean KN, Kraus JP , Biochemistry 2001;40:10625-10633.: Regulation of human cystathionine beta-synthase by S-adenosyl-L-methionine: evidence for two catalytically active conformations involving an autoinhibitory domain in the C-terminal region. PUBMED:11524006
3. Zhang R, Evans G, Rotella FJ, Westbrook EM, Beno D, Huberman E, Joachimiak A, Collart FR; , Biochemistry 1999;38:4691-4700.: Characteristics and crystal structure of bacterial inosine-5'-monophosphate dehydrogenase. PUBMED:10200156
4. Scott JW, Hawley SA, Green KA, Anis M, Stewart G, Scullion GA, Norman DG, Hardie DG; , J Clin Invest 2004;113:274-284.: CBS domains form energy-sensing modules whose binding of adenosine ligands is disrupted by disease mutations. PUBMED:14722619
5. Bateman A; , Trends Biochem Sci 1997;22:12-13.: The structure of a domain common to archaebacteria and the homocystinuria disease protein. PUBMED:9020585
6. Kemp BE; , J Clin Invest 2004;113:182-184.: Bateman domains and adenosine derivatives form a binding contract. PUBMED:14722609

Interpro entry IPR000644

CBS (cystathionine-beta-synthase) domains are small intracellular modules, mostly found in two or four copies within a protein, that occur in several different proteins in all kingdoms of life. Tandem pairs of CBS domains can act as binding domains for adenosine derivatives and may regulate the activity of attached enzymatic or other domains PUBMED:14722619. In some cases, CBS domains may act as sensors of cellular energy status by being activated by AMP and inhibited by ATP PUBMED:14722619. In chloride ion channels, the CBS domains have been implicated in intracellular targeting and trafficking, as well as in protein-protein interactions, but results vary with different channels: in the CLC-5 channel, the CBS domain was shown to be required for trafficking PUBMED:14521953, while in the CLC-1 channel, the CBS domain was shown to be critical for channel function, but not necessary for trafficking PUBMED:14718533.

Figure 2.5.3 View of the CBS domain pair Family homepage. This contains a description of the function of the family and cross-links to other databases. The links on the left allow the user to view the alignments both seed and full, the domain organization, and other family-specific pages as described in the main text. There is also information on when, how, and by whom the family was built.



Family: *CBS* (PF00571)

Domain organisation

Below is a listing of the unique domain organisations or architectures in which this domain is found. More...

There are 1297 sequences with the following architecture: CBS
UMP3_ARATH [arabidopsis thaliana (mouse-ear cress)] protein at5g10860, mitochondrial precursor (206 residues)
Show all sequences with this architecture.

There are 849 sequences with the following architecture: DUF21, CBS, CorC_HlyC
Y260_SYNY3 [synechocystis sp. (strain pcc 6803)] upf0053 protein sll0260 (448 residues)
Show all sequences with this architecture.

There are 633 sequences with the following architecture: IMPDH, CBS
IMDH1_HUMAN [homo sapiens (human)] inosine-5'-monophosphate dehydrogenase 1 (ec 1.1.1.205) (impdehydrogenase 1) (impdh-i) (impd 1) (514 residues)
Show all sequences with this architecture.

There are 437 sequences with the following architecture: Voltage_CLC, CBS
CLCH_TORCA [torpedo californica (pacific electric ray)] chloride channel protein (clc-0) (810 residues)
Show all sequences with this architecture.

There are 402 sequences with the following architecture: MgtE_N, CBS, MgtE
Q8PWC9_METMA [methanosarcina mazei (methanosarcina frisia)] mg2+ transporter mgte (443 residues)
Show all sequences with this architecture.

There are 375 sequences with the following architecture: SIS, CBS
GUTQ_ECOLI [escherichia coli] protein gutq (321 residues)
Show all sequences with this architecture.

Figure 2.5.4 The Pfam view of domain organization within proteins of the CBS (cystathionine-β-synthase) domain family. Pfam-A domains are shown as large-colored boxes, with Pfam-B families as smaller-striped boxes.

View domain organization for a Pfam family

4. View the architecture or domain organization of a family by clicking on the Domain organization tab on the left-hand menu bar on the Family page.

   This page gives a full listing of all the combinations of Pfam domains (Families) found on all the sequences in that Pfam family, referred to as Architectures. Figure 2.5.4 shows the list of all domain architectures for the CBS family. Only one example of each architecture is given, though all instances of that combination of domains can be viewed by clicking on ***Show*** *all sequences with this architecture. The different architectures are ordered according to their prevalence.*

View alignments for a Pfam family

5. Access the alignments of each Pfam family by clicking on the Alignments tab on the left-hand menu bar on the Family page.

   Two types of alignment can be viewed, the seed or the full alignment, as described below (see Background Information). Links to three alignment viewers are given: Jalview—a Java applet, written by Michele Clamp (more information on Jalview is available at http://www.jalview.org/), Pfam viewer, and as HTML. It is strongly advised that alignments be viewed through one of these recommended or related alignment viewers. The alignments are available for downloading, and a number of different formatting options are given including Selex, Stockholm, FASTA, and MSF, as well as several gap-rendering options. Very large alignments can be downloaded as gzip-compressed, Stockholm-format files (e.g., as `PF02171.seed.gz`*).*

View family tree for a Pfam family

6. Obtain an overview tree for the sequences in either the seed or the full alignment by clicking on the Trees tab on the left-hand menu bar of the Family page.

   The Trees page displays a UPGMA-based phylogenetic tree generated using Quicktree (http://www.sanger.ac.uk/Software/analysis/quicktree/), for the given alignment—Seed or Full. These trees give only a rough guide to the sequences that cluster together. For a more accurate and detailed phylogenetic analysis the user should rebuild the tree using more sophisticated software (UNIT 6.1).

View curation and models for a Pfam family

7. View the full curation information on the family and how it was built by clicking on the Curation & models tab in the left-hand menu bar of the Family page.

   The Curation & models tab brings up a page with full information about how the family was curated, including Seed source, domain Type, Author, and statistics on the numbers in seed and full, length of domain, alignment identity, and coverage. The HMM information lists: the HMM build commands, the Gathering, Trusted and Noise cut-offs, Model length, Family (HMM) version, and a link to a visualization of the alignment as an HMM logo. The HMMs of either the ls or fs models (see Guidelines for Understanding Results) may be downloaded. See the APPENDIX at the end of this unit for a more thorough explanation of the flatfile data.

View species distribution for a Pfam family

8. Obtain an overview of phyla within which a domain is found, in the form of a species tree, by clicking on the Species tab on the left-hand menu bar of the Family page (Fig. 2.5.5).

   The Species tab brings up a distribution tree showing the occurrences of this domain across different species for all the sequences in the full alignment. When alignments are relatively small the representation of the tree is interactive, and the user can expand or collapse the tree and view details on it in various ways by using the Tree controls tool box. Species represented in the Seed are, by default, highlighted in blue, but this can be turned off. By default, summary boxes giving, in pink, the numbers of species represented at that taxonomic level, in green, the numbers of sequences found in those species at that level, and, in blue, the number of domains in that species, are displayed, but can also be turned off. Species trees of large families are not interactive and the very largest trees cannot be shown.

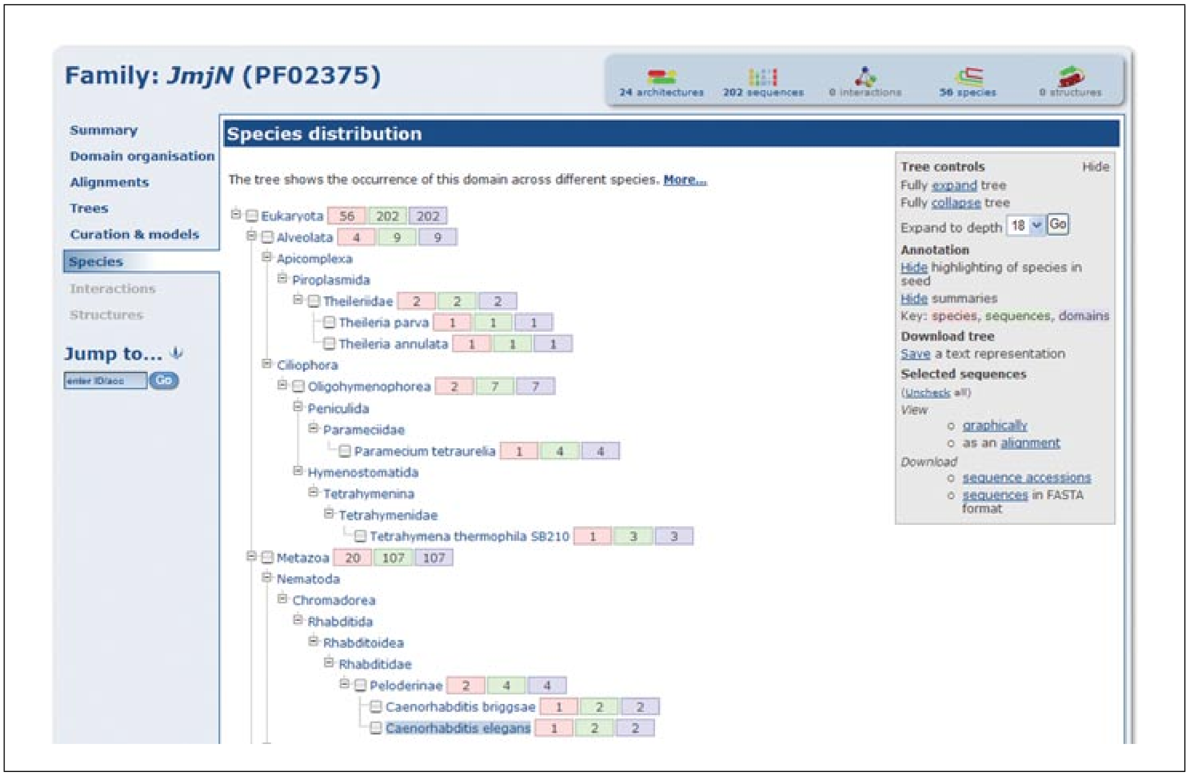
   Sequences or groups of species-specific sequences, as selected by clicking in the checkbox adjacent to the species required, can be viewed as domain graphics or as alignments, and these same selections can be downloaded as lists of sequence accession numbers, alignments, or formatted FASTA sequence.

   Further uses of taxonomy data in Pfam are discussed below (see Suggestions for Further Analysis).

View structures for a Pfam family

9. Obtain a list of all the known protein structures belonging to the family by clicking on the Structures tab on the left-hand menu bar of the Family page.

   For those sequences which have a structure in the Protein Data Bank, Pfam uses the mapping between UniProt, PDB, and Pfam coordinate systems from the MSD group to map Pfam domains onto UniProt three-dimensional structures. The table shows the UniProt sequences that have known structures and the positions within the sequence where these are found. Options to view the structure with Jmol, AstexViewer, or SPICE are given.



The tree shows the occurrence of this domain across different species. More...

Figure 2.5.5 The species distribution of the JmjN domain family. Nodes within the tree are clickable and can be expanded or collapsed. The numbers in brackets represent the number of proteins containing the domain in the respective level of the tree as described in the main text.

Additional analysis

Finding proteins with similar domain organization

10. Ask questions on the presence or absence of domains on particular proteins by clicking on the SEARCH field at the top of the homepage and selecting the Domain architecture tab on the left-hand menu bar (see Fig. 2.5.1).

A menu-driven interface allows retrieval of, for example, all proteins with an Fz, kringle, and Trypsin domain but not an I-set domain, by specifying, from the menus, one or more domains that must appear and one or more domains that must not appear in the resulting architecture.

Finding domains specific to a particular organism

11. Find Pfam families specific to a group of organisms by clicking on the SEARCH field at the top of the homepage and selecting the Taxonomy tab on the left-hand menu bar.

This page allows queries using both simple and complex logic using Boolean AND, OR, *and* NOT *operators. For example, as part of a screen for possible drug targets unique to the malaria parasite, the user might wish to identify all Pfam domains present in Plasmodium falciparum but not in the vertebrate host. The taxonomic query* Plasmodium falciparum AND NOT Vertebrata *will return 43 Pfam domains. Drop-down lists help define the taxonomic level needed.*

Finding families with similar functions

12. Find families with similar functions by clicking on the SEARCH field at the top of the homepage and selecting the Functional similarity tab and pasting in the Pfam family name or number.

This function uses the Gene Ontology terms related to each Pfam family, if known, to search for other families with similar GO terms.

RUNNING Pfam/HMMER LOCALLY

ALTERNATE PROTOCOL 1

To analyze large numbers of protein sequences against Pfam it is preferable to run these searches locally. First, the user needs to download the Pfam release files from the FTP site (linked from the FTP field on the top of homepage). The necessary files are described in detail in the *APPENDIX* at the end of this unit. In order to run the profile HMM searches, the user also needs to download and install the HMMER suite of software (version 2.3.2) written by Sean Eddy, available at *http://hmmer.janelia.org*. HMMER is a software package used by Pfam for building and searching profile HMMs.

The tool used to search query sequences against a profile HMM library is hmmpfam. The format of search commands is:

```
hmmpfam --cut_ga <hmm database> <fasta sequence file>
```

The `--cut_ga` option specifies that the search will use the thresholds curated by the Pfam team as contained in the HMM library. Users may wish to set their own thresholds to view less significant matches. Please see the HMMER documentation for more information on this.

Pfam distributes two HMM libraries—one for finding matches to whole domains, and one for detecting fragment matches. These two are described in more detail in the Guidelines for Understanding Results section. Depending on the user's needs, they may wish to only search one of these libraries or both, and combine the results.

USING Pfam PROFILE HMMs TO FIND DOMAINS IN GENOMIC SEQUENCE

ALTERNATE PROTOCOL 2

A considerable amount of sequence data is released into the public domain as raw genomic sequence. Analysis of this data is hampered by the presence of frameshifting sequencing errors. In addition, the analysis of eukaryotic genomic data is further complicated by the presence of introns. The Wise2 algorithm [developed by Birney et al. (2004); see *http://www.ebi.ac.uk/Wise2/*] compensates for both frame-shifts and introns, incorporates a gene-prediction algorithm, and allows genomic DNA to be compared directly with a protein profile HMM.

From the homepage choose the SEARCH page from the top field and select the DNA sequence tab in the left-hand menu bar and paste in up to 80 kb of ungapped DNA sequence. Because these searches can be time-consuming they are run offline and the results are e-mailed back to the user.

GUIDELINES FOR UNDERSTANDING RESULTS

E-Values

Pfam is based on hidden Markov model (HMM) searches, as provided by the HMMER2 package. In HMMER2, as in BLAST, expectation values, or *E*-values, are calculated as the number of hits that would be expected to have a score equal to or better than a chosen value by chance alone. A well-chosen *E*-value is much <1 (the authors recommend 0.001).

There is a distinction between sequence scores and domain scores. When the user sets the *E*-value cut-off for a Pfam server search, they are setting the cut-off score for the whole sequence rather than for an individual domain. Should the query sequence have a sequence score with an *E*-value less than the cut-off, all individual domains in the query will be reported, even though some of them may have dubious scores individually. This is because in multidomain proteins there are often eroded, weakly matching domains.

Since the Pfam server reports the domain scores and *E*-values, the user may see individual domains with *E*-values worse (i.e., higher) than the specified cut-off.

Bit Scores

The *E*-values described above are based on the raw HMMER scores, or bit scores (see below). In the formalism used by HMMs (essentially a Bayesian statistical formalism), the HMM of the domain is considered as a model which “produces” sequences with a certain probability, different for each different sequence. For each possible sequence, these probabilities are very low, but they are compared with the probability that this sequence was produced by a “random” (or NULL) model. The log-base2 of the ratio of these two likelihoods is reported as the log-odds ratio, commonly called the bit score. In theory, a positive bit score represents a significant match, but in practice, more conservative cut-offs are used, as the size of the database means hits are seen with low positive bit scores occasionally by chance.

Cut-off thresholds

The two scores used for determining the significance of the match between a sequence and an HMM are the *E*-value and the bit score. By default, Web searches use the Pfam-defined Gathering threshold as the cut-off. This cut-off is the same as that used for determining inclusion of sequences into a Pfam full alignment.

These cut-offs are curated to maximize selectivity, such that curators believe there are no false positives falling above these thresholds. In some cases, this may lead to reduced sensitivity. The Web site allows the user to fine-tune the sensitivity by setting their own *E*-value, in order perhaps to find more distantly related matches. On a cautionary note, matches with *E*-values >0.01 should be analyzed with extreme care, as there is a good chance that the domain identified will represent a false match.

Difference between global “ls” and local/fragment “fs” searching

HMMER is most sensitive at identifying complete domains. Its preferred algorithm is a “profile alignment” algorithm that is neither fully global nor fully local, but instead looks for an alignment that is global with respect to the model and multiply local with respect to the sequence—i.e., it looks for one or more complete domains within a query sequence.

Fragments of domains do occur, especially in truncated sequences and translated ESTs, or because of insertions of new domains into existing ones. Each Pfam family consists of two HMMs, one built to find complete domains (ls mode) and the other built to match fragments of domains (fs mode). Having both of these allows Pfam searching to be comprehensive, but at the cost of having to carry out twice as many HMM searches. The most sensitive way to search Pfam is to combine the results from both the ls- and the fs-mode HMMs. However, if the user knows that the query sequence is likely to represent only part of a domain (e.g., ESTs), then it may be more appropriate to search using only the fs option. Alternatively, should the interest be in searching for only complete domains, then using the ls option alone might be more suitable.

The Web site default search is against both the ls and fs HMMs, but options are available for choosing one or the other.

COMMENTARY

Background Information

For the purposes of using the Web site and/or downloading files for local searching, the key elements stored in relation to a Pfam family are the annotations and the multiple sequence alignments both seed and full. The underlying sequence database is called *pfamseq* and is composed of all UniProt sections: SWISS-PROT and SP-TrEMBL (Wu et al., 2006).

For each Pfam-A family the seed alignment contains a manually curated set of representative members from which a profile HMM model (Sonnhammer et al., 1998) is built using the HMMER software package (version 2.3.2; see Alternate Protocol 2). This profile HMM is subsequently used to search *pfamseq* for all detectable matching sequences. The full alignment will then contain all those sequences with an *E*-value falling above a certain curated threshold (see Guidelines for Understanding Results). Full alignments can be very large, with the top 20 families now each containing >20,000 sequences. As an example, the largest full alignment in Pfam, for the HIV GP120 glycoprotein, has >70,000 members, even though the seed alignment from which it was generated consists of just 24 representative members.

Pfam is designed to be both accurate and comprehensive (Finn et al., 2008). To achieve this, the families within it are of two types, referred to as Pfam-A and Pfam-B (Bateman et al., 2002). The accuracy is provided by the high-quality manually annotated Pfam-A families. The comprehensiveness is achieved from the inclusion of Pfam-B families; these are generated automatically from the PRODOM database (Corpet et al., 2000) at each release, see *APPENDIX* at the end of this unit. A more recent level of classification within Pfam is by clans. A clan is a collection of families with similar or related function. Clans are identified on the basis of common sequence or structure similarity and may allow some inference of function between families.

At the time of writing, the latest release of Pfam (release 22.0) contains 9381 Pfam-A families. For coverage, 74% of all sequences in SWISS-PROT 51.7 and SP-TrEMBL 34.7 contain at least one match to Pfam-A, and these families cover ~50% of all residues in the protein databases. The Pfam-B supplement contains another ~182,000 small un-annotated protein families of lower quality. There are also now 283 clans available.

Genome coverage

An important goal of Pfam is to enable rapid automatic classification of predicted proteins into protein domain families. Pfam is used around the world as an aid to genomic annotation. Although Pfam coverage across the sequence databases is high (74%), these databases are biased towards certain protein families and organisms typically reflecting the bias in the underlying sequence databases. A more useful measure of coverage is the fraction of protein sequences in whole genome sequencing projects that are annotated by Pfam. The authors' analyses of a variety of genomes show that Pfam matches between 50% and 90% of the proteins in the genomes examined. Table 2.5.1 shows some examples of genome sequence coverage by Pfam.

Linking to Pfam

Pfam is maintained by a consortium of researchers based in Cambridge (U.K.), Stockholm (Sweden), and Ashburn (U.S.A.). Web authors are encouraged to make links only to stable Pfam homepages, as detailed below.

Homepages

1. *http://pfam.sanger.ac.uk/* (U.K.)
2. *http://pfam.sbc.su.se* (Sweden)
3. *http://pfam.janelia.org* (U.S.)

Table 2.5.1 Five Examples of Pfam (Version 22.0) Genome Coverage

Genome	No. proteins in Pfam	Sequence coverage	No. families in Pfam
M. jannashii	1782	73%	942
E. coli	4604	83%	2070
S. cerevisiae	5778	73%	2166
C. elegans	22433	63%	2651
H. sapiens	37285	67%	3506

Family pages

To link to the family page use either a Pfam accession number or identifier as follows:

1. *http://pfam.sbc.su.se/family?entry = PF00571* (links to accession – Sweden)

2. *http://pfam.sbc.su.sefamily?entry = CBS* (links to id – Sweden)

To link to the different Pfam mirror sites change the root URL.

Protein pages

To link to a page describing an individual protein use the following links:

1. *http://pfam.sanger.ac.uk/protein?entry = P15498* (UniProt Accession)

2. *http://pfam.sanger.ac.uk/protein?entry = VAV_HUMAN* (UniProt Identifier)

Feedback, mailing list, and RSS feed

Pfam encourages feedback from users, especially where interesting domains or families have not been included. Suggestions can be e-mailed to the Pfam curation team at *pfam-help@sanger.ac.uk* or via the blue Add annotation button found on the Family pages (see Fig. 2.5.3).

Users may also wish to join the Pfam mailing list, where new releases of Pfam are announced. To join the list, send an e-mail to *pfamlist-subscribe@sanger.ac.uk*. If you should want to unsubscribe send an e-mail to *pfamlist-unsubscribe@sanger.ac.uk*. Alternatively, you can find out about the latest Pfam news with the RSS feed at *http://pfam.sanger.ac.uk/rss*.

Acknowledgments

The authors would like to acknowledge the hard work of all the members of the Pfam consortium past and present without whom Pfam would not be possible. Finally, we would like to thank all the users of Pfam who have contributed suggestions for improvements over the years.

Literature Cited

Bateman, A., Birney, E., Cerrutti, L., Durbin, R., Etwiller, L., Eddy, S.R., Griffiths-Jones, S., Howe, K.L., Marshall, M., and Sonnhammer, E.L. 2002. The Pfam protein families database. *Nucleic Acids Res.* 30:276-280.

Birney, E., Clamp, M., and Durbin, R. 2004. GeneWise and Genomewise. *Genome Res.* 14:998-995.

Chothia, C. 1992. Proteins. One thousand families for the molecular biologist. *Nature* 357:543-534.

Corpet, F., Servant, F., Gouzy, J., and Kahn, D. 2000. ProDom and ProDom-CG: Tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Res.* 28:267-269.

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.L., and Bateman, A. 2008. The Pfam protein families database. *Nucleic Acids Res.* 36:D281-D288.

Kabsch, W. and Sander, C. 1983. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22:2577-2637.

Mistry, J., Bateman, A., and Finn, R.D. 2007. Predicting active site residue annotations in the Pfam database. *BMC Bioinformatics* 8:298.

Sonnhammer, E.L.L., Eddy, S.R., and Durbin, R. 1997. Pfam: A comprehensive database of protein domain families based on seed alignments. *Proteins* 28:405-420.

Sonnhammer, E.L.L., Eddy, S.R., Birney, E., Bateman, A., and Durbin, R. 1998. Pfam: Multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res.* 26:320-322.

Wu, C.H., Apweiler, R., Bairoch, A., Natale, D.A., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Mazumder, R., O'Donovan, C., Redaschi, N., and Suzek, B. 2006. The Universal Protein Resource (UniProt): An expanding universe of protein information. *Nucleic Acids Res.* 34:D187-D191.

Yooseph, S., Sutton, G., Rusch, D.B., Halpern, A.L., Williamson, S.J., Remington, K., Eisen, J.A., Heidelberg, K.B., Manning, G., Li, W., Jaroszewski, L., Cieplak, P., Miller, C.S., Li, H., Mashiyama, S.T., Joachimiak, M.P., van Belle, C., Chandonia, J.M., Soergel, D.A., Zhai, Y., Natarajan, K., Lee, S., Raphael, B.J., Bafna, V., Friedman, R., Brenner, S.E., Godzik, A., Eisenberg, D., Dixon, J.E., Taylor, S.S., Strausberg, R.L., Frazier, M., and Venter, J.C. 2007. The Sorcerer II Global Ocean Sampling expedition: Expanding the universe of protein families. *PLoS Biol.* 5:e16.

APPENDIX: Pfam DATA AVAILABLE VIA THE WEB SITE

In order to carry out bulk searches or import any Pfam information it will be necessary to access some underlying data. These data are available from the FTP site, the link for which is on the homepage in the top menu bar. The `current_release` directory houses all the files for the latest release. The flatfiles for Pfam-A, both seed and full, and Pfam-B are in Stockholm format—described below. The Pfam-C file gives lists of clans as annotation.

The annotation that is curated for each Pfam-A family provides the core biological information for that family in the form of a structure-function explanation and cross-links to other relevant databases and published literature. It also carries all the nonbiological information on how the family was built and the alignments curated.

Stockholm format

This is a flexible format for files including alignments where the specification is based on four tags, described briefly below:

```
#=GF <feature> <per file annotation>

#=GC <feature> <per column annotation>

#=GS <sequence name> <feature> <per sequence annotation>

#=GR <sequence name> <feature> <per sequence AND per
column annotation>
```

This format allows for the addition of sequence descriptors, such as the SS lines for secondary-structure and AS lines for active site mark-up. More information about the Stockholm format can be found at *http://sonnhammer.sbc.su.se/Stockholm.html*.

Figure 2.5.6 shows the flatfile for Family 7kD_DNA_binding in Stockholm format.

Useful files

a. The `rel_notes.txt` file carries all relevant information on how to use Pfam, as well as listing the major changes from the last to this release.

b. The `userman.txt` contains a detailed description of the annotation fields—see Background Information.

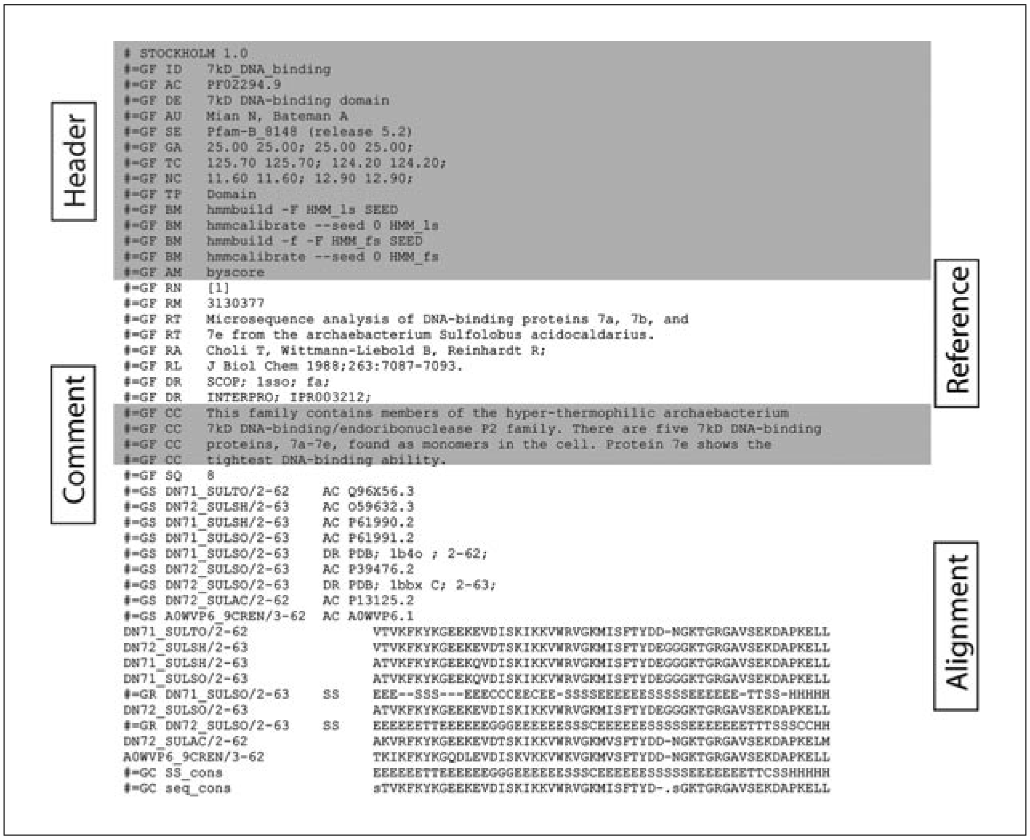
c. The `diff` file contains a list of Pfam-entries that have changed since the last release, and will be useful if the goal is to update a previous analysis.

d. The `Pfam-A.full` and `Pfam-A.seed` are the two files of the multiple alignments for the Full, or the Seed, and both carry the family annotation.

e. The `Pfam-B` file contains all the families in the database that are created automatically, with their identifier and their multiple sequence alignment.

f. `Pfamseq` is the underlying sequence database and is provided in FASTA format (*APPENDIX 1B*).

g. The `swisspfam` file contains crude ASCII-type tables showing the positions of domains on a protein if present, and is probably simpler to consult for a quick check on presence or absence of domains on a protein than the `Pfam-A.full` file.



```
# STOCKHOLM 1.0
#=GF ID   7kD_DNA_binding
#=GF AC   PF02294.9
#=GF DE   7kD DNA-binding domain
#=GF AU   Mian N, Bateman A
#=GF SE   Pfam-B_8148 (release 5.2)
#=GF GA   25.00 25.00; 25.00 25.00;
#=GF TC   125.70 125.70; 124.20 124.20;
#=GF NC   11.60 11.60; 12.90 12.90;
#=GF TP   Domain
#=GF BM   hmmbuild -F HMM_ls SEED
#=GF BM   hmmcalibrate --seed 0 HMM_ls
#=GF BM   hmmbuild -f -F HMM_fs SEED
#=GF BM   hmmcalibrate --seed 0 HMM_fs
#=GF AM   byscore
#=GF RN   [1]
#=GF RM   3130377
#=GF RT   Microsequence analysis of DNA-binding proteins 7a, 7b, and
#=GF RT   7e from the archaebacterium Sulfolobus acidocaldarius.
#=GF RA   Choli T, Wittmann-Liebold B, Reinhardt R;
#=GF RL   J Biol Chem 1988;263:7087-7093.
#=GF DR   SCOP; 1sso; fa;
#=GF DR   INTERPRO; IPR003212;
#=GF CC   This family contains members of the hyper-thermophilic archaebacterium
#=GF CC   7kD DNA-binding/endoribonuclease P2 family. There are five 7kD DNA-binding
#=GF CC   proteins, 7a-7e, found as monomers in the cell. Protein 7e shows the
#=GF CC   tightest DNA-binding ability.
#=GF SQ   8
#=GS DN71_SULTO/2-62    AC Q96X56.3
#=GS DN72_SULSH/2-63    AC O59632.3
#=GS DN71_SULSH/2-63    AC P61990.2
#=GS DN71_SULSO/2-63    AC P61991.2
#=GS DN71_SULSO/2-63    DR PDB; 1b4o ; 2-62;
#=GS DN72_SULSO/2-63    AC P39476.2
#=GS DN72_SULSO/2-63    DR PDB; 1bbx C; 2-63;
#=GS DN72_SULAC/2-62    AC P13125.2
#=GS A0WVP6_9CREN/3-62  AC A0WVP6.1
DN71_SULTO/2-62               VTVKFKYKGEEKEVDISKIKKVWRVGKMISFTYDD-NGKTGRGAVSEKDAPKELL
DN72_SULSH/2-63               VTVKFKYKGEEKEVDTSKIKKVWRVGKMISFTYDEGGGKTGRGAVSEKDAPKELL
DN71_SULSH/2-63               ATVKFKYKGEEKQVDISKIKKVWRVGKMISFTYDEGGGKTGRGAVSEKDAPKELL
DN71_SULSO/2-63               ATVKFKYKGEEKQVDISKIKKVWRVGKMISFTYDEGGGKTGRGAVSEKDAPKELL
#=GR DN71_SULSO/2-63    SS    EEE--SSS---EEECCCEECEE-SSSSEEEEEESSSSSEEEEEE-TTSS-HHHHH
DN72_SULSO/2-63               ATVKFKYKGEEKEVDISKIKKVWRVGKMISFTYDEGGGKTGRGAVSEKDAPKELL
#=GR DN72_SULSO/2-63    SS    EEEEEETTEEEEEEGGGEEEEEEESSSCEEEEEESSSSSEEEEEEEETTTSSSCCHH
DN72_SULAC/2-62               AKVRFKYKGEEKEVDTSKIKKVWRVGKMVSFTYDD-NGKTGRGAVSEKDAPKELM
A0WVP6_9CREN/3-62             TKIKFKYKGQDLEVDISKVKKVWKVGKMVSFTYDD-NGKTGRGAVSEKDAPKELL
#=GC SS_cons                  EEEEEETTEEEEEEGGGEEEEEEESSSCEEEEEESSSSSEEEEEEEETTCSSHHHHH
#=GC seq_cons                 sTVKFKYKGEEKEVDISKIKKVWRVGKMISFTYD-.sGKTGRGAVSEKDAPKELL
```

Figure 2.5.6 The entry in the `Pfam-A.full` file for the family 7kD_DNA_binding, in Stockholm format. The different sections (header, references, comment, and alignment) are labeled. Fields are described in detail in Table 2.5.2 and the main text.

h. The `Pfam_ls` and `Pfam_fs` are the global and local/fragment files of the profile-HMMs. `Pfam_ls` has the profile-HMMs for finding complete domain matches, whereas `Pfam_fs` has the HMMS for finding incomplete domain matches to a sequence.

i. The `Pfam-C` file contains all the information on the Pfam clans, the collections of related families.

Pfam-A flatfiles

The Pfam-A flatfiles include annotation and alignments in Stockholm format for each family. Each family is divided into an annotation and alignment section. Fields in these sections are described in Table 2.5.2.

Annotation

Header section

The header division of the annotation file contains compulsory fields describing the family. These data include the Pfam accession number and the unique identifying family name, together with a slightly fuller description of the function. The one noncompulsory field in the header section is the previous identification (PI) field—given when a family changes its name.

The type field (TP) is particularly worthy of elaboration. Although the authors try to classify proteins in Pfam into distinct domains within Pfam, this is not always possible. There are now four types of family, defined as follows. The default type is **Family**, which

Table 2.5.2 Description of All Pfam Annotation Fields

Field name	Field tag	Description
Accession number	AC	This field contains one word in the form PF*xxxxx* or CL*xxxx*, where *x* is a digit. The Pfam-A and Pfam-C accession numbers are stable identifiers for each Pfam family and do not change between releases.
Identification	ID	A single short word (<16 characters) that is more meaningful than the accession number. This identifier may change between Pfam releases.
Previous ID(s)	PI	The identifier (ID) of a family may change over time. To keep track of these changes, the PI field is used. This field is optional.
Definition	DE	A one-line description of the Pfam family.
Author	AU	The authors of the Pfam entry.
Type	TP	Describes the family type. Presently, the entry can be one of the following types: Family, Domain, Repeat, or Motif (defined in the main text).

Build method	BM	The build method field contains the HMM building command lines. An example of the BM lines from a single entry are:
		`BM hmmbuild -F HMM_ls SEED BM hmmcalibrate -seed 0 HMM_ls`
		`BM hmmbuild -f -F HMM_fs SEED BM hmmcalibrate -seed 0 HMM_fs.`
		These command lines can be used to rebuild the Pfam HMMs from the seed alignments.

Alignment order	AM	Indicates the order that the ls and fs matches are aligned to the model to give the full alignment (can be one of globalfirst, localfirst, or byscore).
Source of seed	SE	The source suggesting that the seed members might belong to a family. This field can indicate a Pfam-B accession number, another database, or that the seed came from a reference cited in the DESC file.
Gathering method	GA	GA lines are the thresholds used in the hmmsearch command line. This line is described elsewhere in the section on Pfam thresholds. An example of this line is `GA 25.0 3.2; 24.0 0.8.`
Threshold cutoff	TC	Gives scores of lowest-scoring real matches above the GA threshold. An example TC line is `TC 26.2 3.2; 24.6 1.2.`
Noise cutoff	NC	Gives scores of highest scoring noise. An example NC line is `NC 24.0 23.9; 21.5 16.3.`
Nested domain	NE	Gives the accession number of a Pfam family that may be nested within the alignment. The nested family is allowed to overlap with this one.
Nested location	NL	This line gives a location of a nested domain in the SEED alignment.

Database reference	DR	Reference to an external database. Pfam carries links to a variety of databases, this information is found in DR lines. The format is: `<DR Database>; <Identifier>.`
		For example:
		`DR PFAMB; PB000001;`
		`DR PRINTS; PR00012;`
		`DR URL; http://bmerc.bu.edu/projects/wdrepeat;`
		For SCOP links, a third field is added indicating the level of placement in the SCOP hierarchy—e.g., `DR SCOP; 1pii; fa;.`
		For PDB links, the second field contains the PDB identifier and chain identifier if present. The third and fourth fields contain the start and end points within the PDB entry—e.g., `DR PDB; 2nad A; 123; 332;.`

continued

Table 2.5.2 Description of All Pfam Annotation Fields, *continued*

Field name	Field tag	Description
Database comment	DC	Comment for a database reference.
Reference comment	RC	Comment for a literature reference.
Reference number	RN	Digit in square brackets representing reference number for citations in Pfam comments field.
Reference PubMed	RM	The PubMed PMID number.
Reference title	RT	Title of the paper.
Reference author(s)	RA	A list of the authors responsible for the paper. An example of the format is: `RA Bateman A, Eddy SR, Mesyanzhinov VV`.
Reference location	RL	The reference location is in the following format: `RL <Journal abbreviation> <year>;<volume:page-page>`.
Comment	CC	A manually annotated description of the family.

simply implies that the members are related. A **Domain** is defined as an autonomous structural unit, or a reusable sequence unit that may be found in multiple protein contexts or architectures. In contrast, a **Repeat** is not normally stable in isolation, but rather is required in multiple tandem repeats to form a globular domain or extended structure. **Motifs** generally describe shorter sequence units found outside of globular domains. Pfam release 22.0 contains 6334 families, 2423 domains, 146 repeats, and 54 motifs.

Reference section

The reference section contains cross-links to other databases and literature references. The literature references are not intended to be comprehensive, but to provide the user with a good starting point for understanding the family.

Comment section

Many families contain a textual annotation that describes the structure and function of members of that family and hence, by analogy, of the whole family.

Alignment: Aligned sequences

Immediately following, underneath the Comment lines in the flatfiles, come the sequences, as shown in Figure 2.5.6.

Alignment Features

There are at least three different types of nonsequence features tagged with # = GR, which are the SS, the AS, and the PDB structure lines.

The SS line initiator indicates the known secondary structure elements of the sequence, derived from DSSP (Kabsch and Sander, 1983). The DSSP codes for the different elements are given in Table 2.5.3.

The AS line initiator indicates the known active site residues within the family. These are derived from the ACT_SITE-feature-table lines from the SWISS-PROT entry for that protein (see Mistry et al., 2007). Multiple alignments with added mark up clearly show whether active site residues are conserved across all members of a family. The most frequently occurring residues found at active sites, as tagged in SWISS-PROT, are C, D, E, H, K, R, S, and Y.

The PDB structures for any sequence, if known, are given.

Table 2.5.3 Definitions of DSSP Codes

DSSP code	DSSP definition
C	Random coil
H	α-helix
G	3_{10} helix
I	5 helix (π-helix)
E	H-bonded β-strand (extended strand)
B	Residue in isolated β-bridge
T	H-bonded turn (3-, 4-, or 5-turn)
S	Bend (five-residue bend centered at residue *i*)

Pfam-B Flatfiles

These contain just the Pfam identifier and the aligned sequences together with any known nonsequence features. Pfam-B is an automatically generated supplement to Pfam-A that provides completeness in terms of coverage. It has also provided a useful resource for generating new Pfam-A families, to the extent that it is now the major source of such new families. Pfam-B is constructed from the PRODOM database (Corpet et al., 2000) of protein domain families. PRODOM is a high-quality, automatically generated database of protein families, and is constructed over the same underlying sequence database as the rest of Pfam. In essence, Pfam-B comprises the parts of PRODOM not already covered by Pfam-A. Pfam-B is newly generated at each release.

Clan Flatfiles

The data on clans is in the form of a numerical list of the clans. The information on each clan is presented against annotation headings, as in Table 2.5.2, rather than in Stockholm format, with the family members of the clan being listed at the bottom against an MB field-tag.

Swisspfam Files

Data on domains present on a protein is also stored in relation to the SWISS-PROT identifier, in the `swisspfam` file. This format provides an ASCII view of protein domain composition.