Using CATH-Gene3D to Analyze the Sequence, Structure, and Function of Proteins

Ian Sillitoe,[1] Tony Lewis,[1] and Christine Orengo[1]

[1]University College London, London, United Kingdom

The CATH database is a classification of protein structures found in the Protein Data Bank (PDB). Protein structures are chopped into individual units of structural domains, and these domains are grouped together into superfamilies if there is sufficient evidence that they have diverged from a common ancestor during the process of evolution. A sister resource, Gene3D, extends this information by scanning sequence profiles of these CATH domain superfamilies against many millions of known proteins to identify related sequences. Thus the combined CATH-Gene3D resource provides confident predictions of the likely structural fold, domain organisation, and evolutionary relatives of these proteins. In addition, this resource incorporates annotations from a large number of external databases such as known enzyme active sites, GO molecular functions, physical interactions, and mutations. This unit details how to access and understand the information contained within the CATH-Gene3D Web pages, the downloadable data files, and the remotely accessible Web services. © 2015 by John Wiley & Sons, Inc.

Keywords: protein structure • protein domain • protein classification • functional family • superfamily

How to cite this article:
Sillitoe, I., Lewis, T. and Orengo, C. 2015. Using CATH-Gene3D to analyze the sequence, structure, and function of proteins. *Curr. Protoc. Bioinform.* 50:1.28.1-1.28.21.
doi: 10.1002/0471250953.bi0128s50

INTRODUCTION

There are already many tens of millions of known protein sequences, and the continuing advances of DNA sequencing techniques suggest that this number will continue to grow rapidly. A huge challenge lies in developing tools that help us to add meaning to this data both quickly and accurately. Ideally, we would like to be able to understand the function of these novel protein sequences and predict their role in a number of biological processes. Since protein function is intrinsically linked to the protein's physical shape, a detailed understanding of three-dimensional structure is critical to understanding and predicting protein function. In addition, similarities in structure can be recognised even when the amino acid sequence has diverged beyond recognition. Thus, protein structure can also prove a valuable tool when probing for very remote evolutionary relationships.

The CATH database provides an expertly curated set of protein structural domains (Sillitoe et al., 2015). CATH domains are discrete, structurally compact units that often exist in many different proteins and can be thought to have semi-independent evolutionary paths to the proteins in which they are found. The domains are identified from the three-dimensional coordinates of protein structures submitted to the Protein Data Base (PDB;

Current Protocols in Bioinformatics 1.28.1-1.28.21, June 2015
Published online June 2015 in Wiley Online Library (wileyonlinelibrary.com).
doi: 10.1002/0471250953.bi0128s50

Bernstein et al., 1977), and domains that are related to each other through a common ancestor are grouped into CATH superfamilies. Evolutionary relationships are identified on the basis of sequence, structural, and functional information; and manual verification is used where the levels of similarity reported by automatic analyses are borderline.

The majority of superfamilies are relatively small with highly conserved structures and functions. However, a small number of superfamilies are very highly populated and contain a hugely diverse set of distantly related structures and functions. CATH recognises over 2,600 superfamilies, however >50% of the entire data set of known protein sequences are contained in <100 of these superfamilies.

CATH organises superfamilies further by clustering them into fold groups where they share similar structural features and then architectures where their secondary structures form similar three-dimensional shapes. Finally, architectures are grouped according to whether they are mainly alpha, mainly beta, or alpha-beta. This accounts for the four levels in the CATH classification: **C**lass, **A**rchitecture, **T**opology, and **H**omologous superfamily. The CATH superfamily numbering scheme reflects this hierarchy. For example, the homologous superfamily 1.10.490.10 (globins) belongs to class 1 (mainly alpha), architecture 1.10 (orthogonal bundle), and topology 1.10.490 (globin-like).

While structure is certainly related to function, it is important to note that identifying a structural similarity does not necessarily mean that two domains perform the same function. Since the ultimate intention is to identify functional similarity, it is important to subclassify the highly diverse superfamilies into clusters that only contain functionally similar domains. CATH refers to these "functionally pure" subsets of domains as functional families, or FunFams. Thus, identifying a strong relationship between a new protein sequence and a CATH superfamily can provide useful information about the likely protein structure, domain organisation, and functional annotation of evolutionary relatives. However, identifying a strong relationship with a CATH FunFam provides far more compelling evidence of the new protein's potential function.

This unit describes how to access and understand the information in the CATH database: either searching with a new protein sequence (Basic Protocol 1) or with a new protein structure (Basic Protocol 2). Also described are searching through downloadable data files (see Suggestions for Further Analysis) and through Web services (see Alternate Protocol).

BASIC PROTOCOL 1

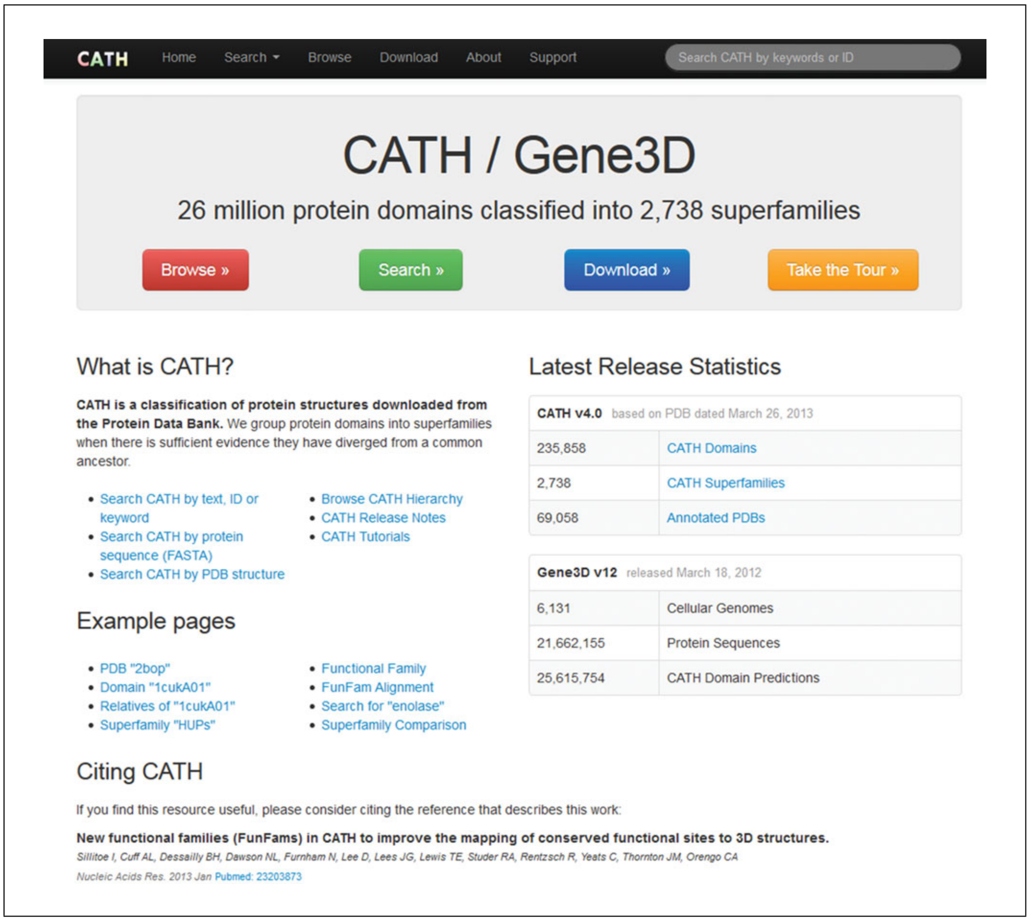
SEARCHING CATH WITH A NEW PROTEIN SEQUENCE

One simple and fast way of finding out information about a new protein is to compare the sequence or structure against a set of known protein domains. If any of our known domains match well against regions of the new protein, then we can use this similarity to infer information. The closer the match, the more confident we can be that this region of the new protein is related to the known domain.

In short, the general strategy of searching CATH is to:

Identify matches to known CATH entities (domains or FunFams)
Display what we already know about these matches
Suggest what is safe to infer about your protein sequence

The most up-to-date documentation for the usage of CATH is found on the Web site. We aim to provide regular updates and improvements to the Web site, so there may be slight differences between the presented figures and the live pages. However, the URLs provided and the core functionality should not be affected.



CATH / Gene3D

26 million protein domains classified into 2,738 superfamilies

What is CATH?

CATH is a classification of protein structures downloaded from the Protein Data Bank. We group protein domains into superfamilies when there is sufficient evidence they have diverged from a common ancestor.

- Search CATH by text, ID or keyword
- Search CATH by protein sequence (FASTA)
- Search CATH by PDB structure
- Browse CATH Hierarchy
- CATH Release Notes
- CATH Tutorials

Latest Release Statistics

CATH v4.0 based on PDB dated March 26, 2013	
235,858	CATH Domains
2,738	CATH Superfamilies
69,058	Annotated PDBs

Gene3D v12 released March 18, 2012	
6,131	Cellular Genomes
21,662,155	Protein Sequences
25,615,754	CATH Domain Predictions

Example pages

- PDB "2bop"
- Domain "1cukA01"
- Relatives of "1cukA01"
- Superfamily "HUPs"
- Functional Family
- FunFam Alignment
- Search for "enolase"
- Superfamily Comparison

Citing CATH

If you find this resource useful, please consider citing the reference that describes this work:

New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures.
Sillitoe I, Cuff AL, Dessailly BH, Dawson NL, Furnham N, Lee D, Lees JG, Lewis TE, Studer RA, Rentzsch R, Yeats C, Thornton JM, Orengo CA
Nucleic Acids Res. 2013 Jan Pubmed: 23203873

Figure 1.28.1 The CATH home page, showing the search bar allowing free text search (top right of all pages) with links to search CATH by sequence (FASTA) or structure (PDB).

Necessary Resources

Hardware

Workstation with network connections

Software

Javascript-enabled Web browser such as Firefox or Internet Explorer

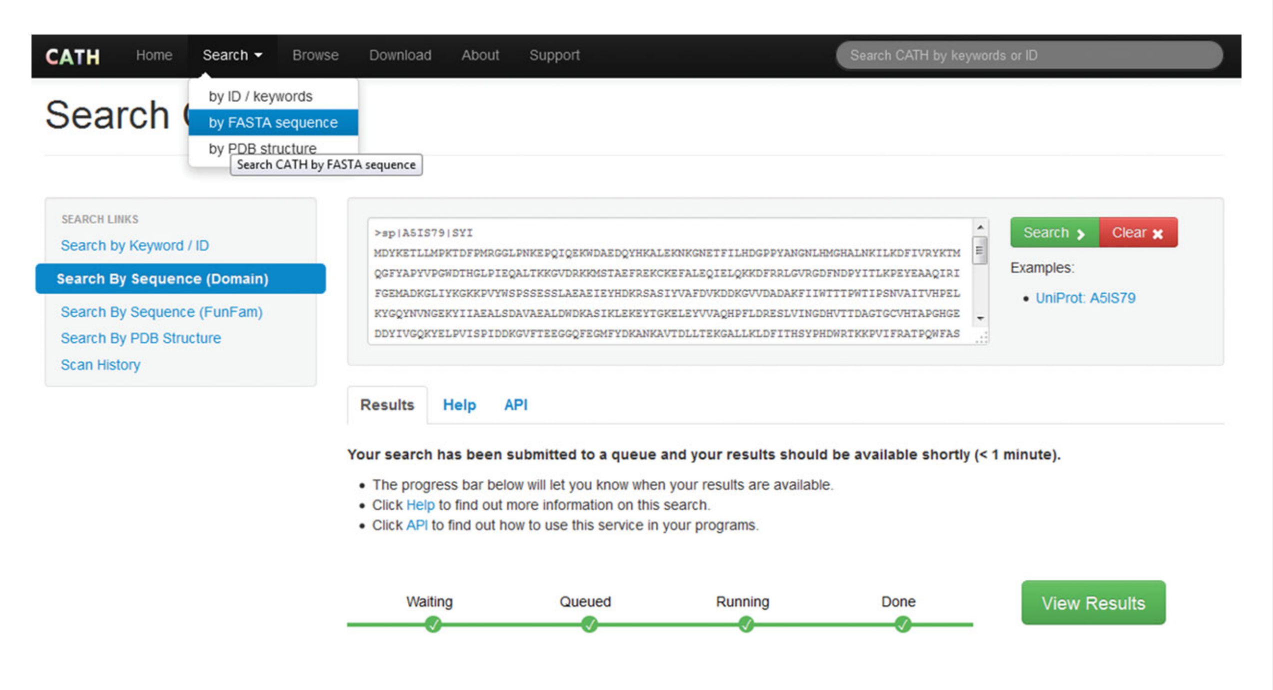
Files

Protein sequence (FASTA format)

1. Direct your browser to the main CATH Web site located at *http://www.cathdb.info* (Fig. 1.28.1).

   The home page provides a number of links to common ways of accessing information in CATH: searching by a biological identifier or keyword, searching by sequence (FASTA), and searching by structure (PDB). It also contains links to example pages to show typical sets of Web pages for particular entities in CATH (domain, superfamily, and FunFam). The top menu bar is constant across the site and provides short links to searching CATH, browsing the hierarchy, downloading data, and finding out more about CATH.

2. Search CATH with a new protein sequence. On the top menu bar of any page, click "Search," then click "by FASTA sequence" (Fig. 1.28.2). There are two different sequence searches available: searching against CATH Domains and searching against CATH FunFams. Both involve scanning a new protein sequence against a CATH



```
>sp|A5IS79|SYI
MDYKETLLMPKTDFPMRGGLPNKEPQIQEKWDAEDQYHKALEKNKGNETFILHDGPPYANGNLHMGHALNKILKDFIVRYKTM
QGFYAPYVPGWDTHGLPIEQALTKKGVDRKKMSTAEFREKCKEFALEQIELQKKDFRRLGVRGDFNDPYITLKPEYEAAQIRI
FGEMADKGLIYKGKKPVYWSPSSESSLAEAEIEYHDKRSASIYVAFDVKDDKGVVDADAKFIIWTTTPWTIPSNVAITVHPEL
KYGQYNVNGEKYIIAEALSDAVAEALDWDKASIKLEKEYTGKELEYVVAQHPFLDRESLVINGDHVTTDAGTGCVHTAPGHGE
DDYIVGQKYELPVISPIDDKGVFTEEGGQFEGMFYDKANKAVTDLLTEKGALLKLDFITHSYPHDWRTKKPVIFRATPQWFAS
```

Your search has been submitted to a queue and your results should be available shortly (< 1 minute).

- The progress bar below will let you know when your results are available.
- Click Help to find out more information on this search.
- Click API to find out how to use this service in your programs.

Figure 1.28.2 Searching CATH by protein sequence.

library (built from CATH domains and CATH FunFams respectively); however, there are important differences:

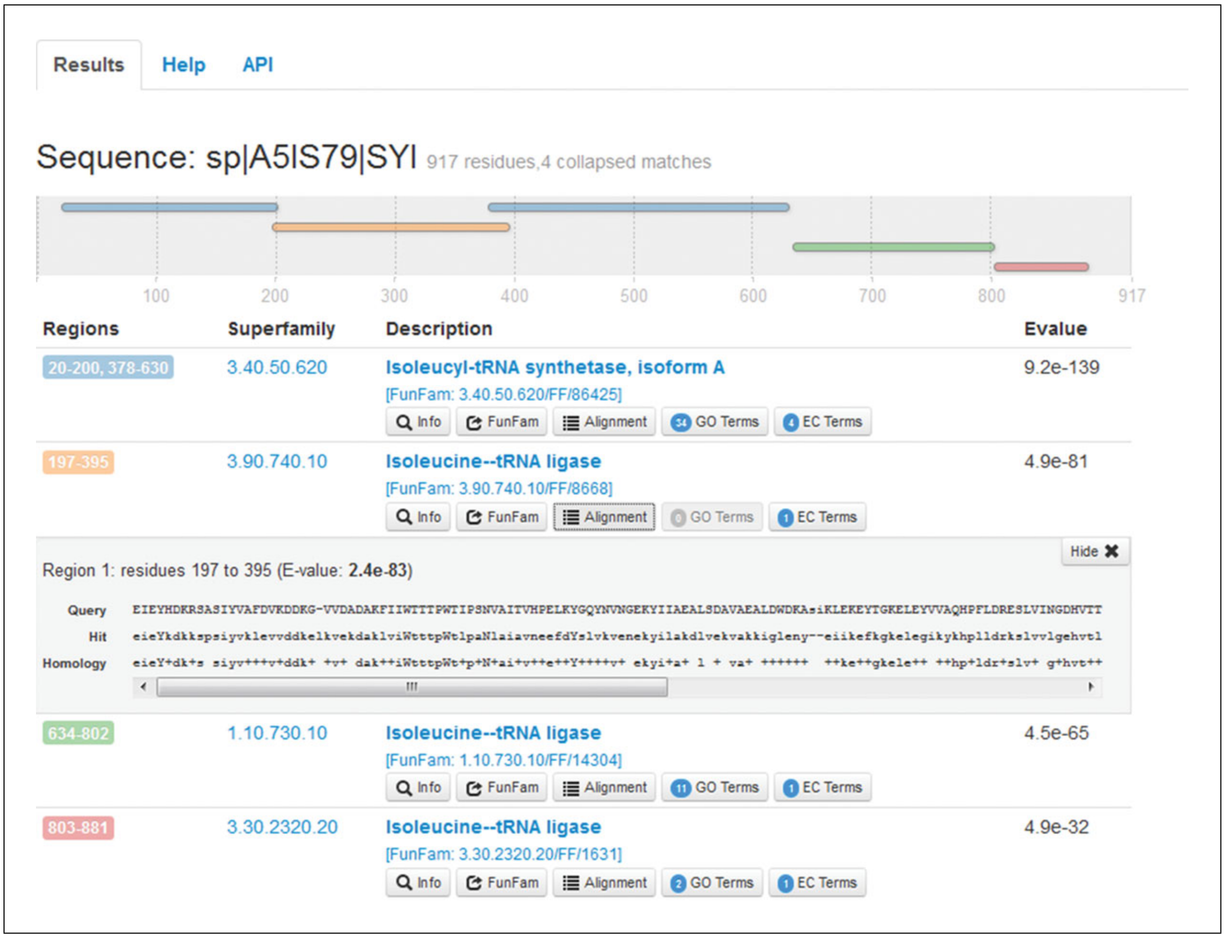
Search by Sequence (Domain) identifies the location of CATH domains that match regions of the query protein sequence. Search by Sequence (FunFam) identifies the location of CATH FunFams that match regions of the query protein sequence. The protein sequences included in a particular FunFam have been deliberately constrained to a particular function; thus, we can be more confident of inferring functional information from FunFam matches rather than Domain matches.

In general, the recommended strategy would be to search a new sequence against CATH FunFams first (as the results are more informative), then search against CATH Domains if no matches are found (as the coverage is wider). The procedure for both scan types is similar; however, the results for the FunFam matches have more detail due to the added functional annotations. As a result, the FunFam scan will be used to illustrate both scan types in the following steps.

a. *Submit*. Paste the query protein sequence into the text box. Then, submit the search by clicking the green button to the right, labelled “Search.”

b. *Monitor*. After submission, a message should indicate that the scan has been added to a queue and that the results will be available shortly. The sequence is being scanned against an HMM library, essentially a collection of sequence-based fingerprints for CATH Domains or CATH FunFams. The progress of the scan is monitored regularly, and the message will change again when the results are available. Typically, these sequence scans take a few seconds, however they can take up to 2 min depending on the length of the sequence and the number of scans in the queue.

c. *Results*. The sequence results provide an overview of the regions of the query protein sequence that have matched entities from CATH (Domains or FunFams). The full set of sequence matches can typically contain hundreds of matches, so the algorithm DomainFinder3 (Yeats et al., 2010) is used to “collapse” these results into the most significant match for each region of query protein (Fig. 1.28.3).



Sequence: sp|A5IS79|SYI 917 residues,4 collapsed matches

Regions	Superfamily	Description	Evalue
20-200, 378-630	3.40.50.620	Isoleucyl-tRNA synthetase, isoform A [FunFam: 3.40.50.620/FF/86425] Info · FunFam · Alignment · 34 GO Terms · 1 EC Terms	9.2e-139
197-395	3.90.740.10	Isoleucine--tRNA ligase [FunFam: 3.90.740.10/FF/8668] Info · FunFam · Alignment · 0 GO Terms · 1 EC Terms	4.9e-81
634-802	1.10.730.10	Isoleucine--tRNA ligase [FunFam: 1.10.730.10/FF/14304] Info · FunFam · Alignment · 11 GO Terms · 1 EC Terms	4.5e-65
803-881	3.30.2320.20	Isoleucine--tRNA ligase [FunFam: 3.30.2320.20/FF/1631] Info · FunFam · Alignment · 2 GO Terms · 1 EC Terms	4.9e-32

Region 1: residues 197 to 395 (E-value: **2.4e-83**)

```
   Query EIEYHDKRSASIYVAFDVKDDKG-VVDADAKFIIWTTTPWTIPSNVAITVHPELKYGQYNVNGEKYIIAEALSDAVAEALDWDKAsiKLEKEYTGKELEYVVAQHPFLDRESLVINGDHVTT
     Hit eieYkdkkspsiyvklevvddkelkvekdaklviWtttpWtlpaNlaiavneefdYslvkvenekyilakdlvekvakkigleny--eiikefkgkelegikykhplldrkslvvlgehvtl
Homology eieY+dk+s siyv+++v+ddk+ +v+ dak++iWtttpWt+p+N+ai+v++e++Y++++v+ ekyi+a+ l + va+ ++++++  ++ke++gkele++ ++hp+ldr+slv+ g+hvt++
```

Figure 1.28.3 Results when searching a protein sequence against CATH FunFams.

The query sequence is represented schematically by the gray region at the top of the page, with a scale to provide the approximate residue numbering. Each matching region is displayed as a colored pill on this gray scale, and a list below the schematic diagram describes each matched CATH entity in more detail. The diagram and the list of results are linked, so moving the mouse over the pill will highlight the relevant match details and vice versa.

Under the description for a particular match, there are a set of buttons that enable the user to interrogate more information about the matching domains in CATH.

Info: provides a quick preview of the FunFam (the number of domains, sequences, functional annotations, etc.)

FunFam: jumps to the page describing the matching FunFam in more detail

Alignment: displays the snippet of the alignment between the query protein sequence and the matched FunFam

GO Terms: displays information on the known functional terms associated with the matching FunFam—based on annotations from the Gene Ontology (GO) database. The number displayed on the button signifies the unique number of GO terms associated with that FunFam.

EC Terms: displays information on known catalytic activity of the matching FunFam—based on annotations from the Enzyme Commission (EC) database. The number displayed on the button signifies the unique number of EC terms associated with that FunFam.

ALTERNATE PROTOCOL

ACCESS CATH SEQUENCE SCAN REMOTELY

The most straightforward method for searching a new protein sequence against the CATH database is through the Web pages as described in Basic Protocol 1. An alternative method for accessing the same underlying information is via the CATH application program interface (API). This allows protein sequences to be searched against CATH without the user having to manually interact with the Web pages (e.g., if a scan needs to be performed repeatedly as part of a larger, automated process). This alternative protocol can be safely skipped if remote access is not required.

The Web pages for these sequence scans have been implemented on top of a simple, representational state transfer (REST) interface to ensure these services can be accessed programmatically as well as through the Web browser. The interface is described in detail on the Web pages by clicking on the API tab of the sequence search tool of interest (*http://www.cathdb.info/search/by_fasta* for Domains and *http://www.cathdb.info/search/by_funfhmmer* for FunFams).

The API tab on the sequence search Web page provides example snippets of code to demonstrate access through basic tools available on Linux (cURL) and a typical programmatic implementation in Perl (Fig. 1.28.4). The procedure for both types of scan is similar and is broken down into three stages. In each case, the "Accept" HTTP header of the request is used to decide the format to return any output, so adding the following to the HTTP headers will ensure the output is formatted as a JSON object:

```
Accept: application/json
```

Further details on each stage are presented below.

Necessary Resources

Hardware

Workstation with network connections

Software

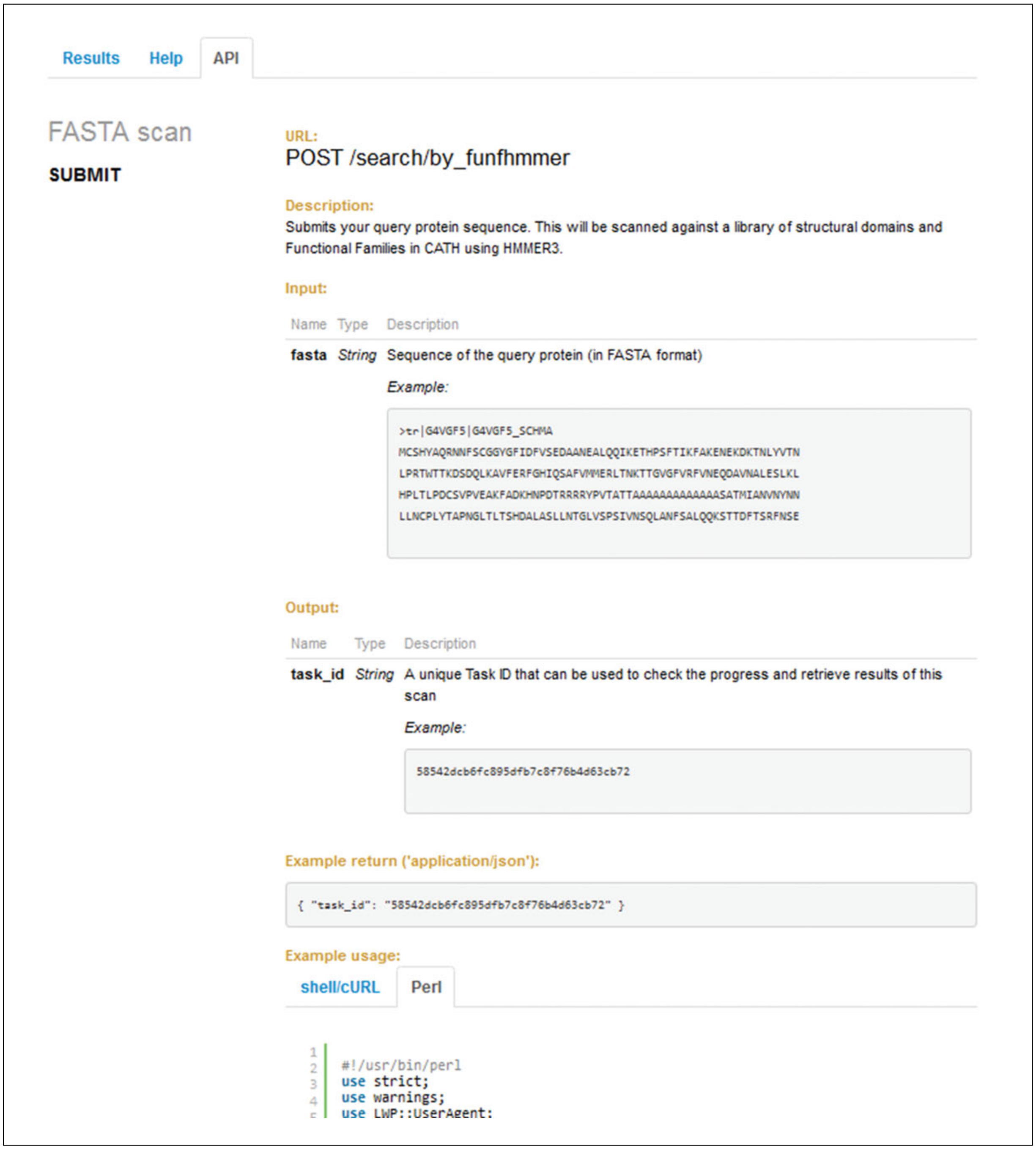
Perl modules (installed via CPAN): `LWP::Simple, JSON::Any`

1. *Submit*. This adds a query protein sequence to a queue to be scanned and returns a unique identifier that can be used in subsequent stages. A search is submitted by issuing a POST request with the query protein sequence sent in a parameter labelled "fasta."

   URL:
   `POST` *`http://www.cathdb.info/search/by_fasta`*
   Input:
   `fasta=<String> query sequence in FASTA format`
   Output:
   `task_id=<String> e.g`
   `"58542dcb6fc895dfb7c8f76b4d63cb72"`
   Examples of working code can be copied and pasted from the API tab on the sequence search Web pages for each stage.

2. *Monitor*. This stage tracks the current status of the scan. The unique identifier returned by the submit stage should be substituted in place of `<task_id>` in the examples below:

   URL:
   `GET` *`http://www.cathdb.info/search/by_fasta/check/<task_id>`*



FASTA scan

SUBMIT

URL:
POST /search/by_funfhmmer

Description:
Submits your query protein sequence. This will be scanned against a library of structural domains and Functional Families in CATH using HMMER3.

Input:

Name	Type	Description
fasta	*String*	Sequence of the query protein (in FASTA format)

Example:

```
>tr|G4VGF5|G4VGF5_SCHMA
MCSHYAQRNNFSCGGYGFIDFVSEDAANEALQQIKETHPSFTIKFAKENEKDKTNLYVTN
LPRTWTTKDSDQLKAVFERFGHIQSAFVMMERLTNKTTGVGFVRFVNEQDAVNALESLKL
HPLTLPDCSVPVEAKFADKHNPDTRRRRYPVTATTAAAAAAAAAAAAASATMIANVNYNN
LLNCPLYTAPNGLTLTSHDALASLLNTGLVSPSIVNSQLANFSALQQKSTTDFTSRFNSE
```

Output:

Name	Type	Description
task_id	*String*	A unique Task ID that can be used to check the progress and retrieve results of this scan

Example:

```
58542dcb6fc895dfb7c8f76b4d63cb72
```

Example return ('application/json'):

```
{ "task_id": "58542dcb6fc895dfb7c8f76b4d63cb72" }
```

Example usage:

```

#!/usr/bin/perl
use strict;
use warnings;
use LWP::UserAgent:
```

Figure 1.28.4 Details of how to use the CATH sequence search remotely are provided in the API section of the Web pages.

Output:

```
success=<Boolean> whether the scan has finished
message=<String> information about the current
  progress
data=<Object> further details about the scan
```

3. *Results*. After the monitor stage has declared that the scan has finished (`success=true`), the results can be retrieved.

URL:

GET *http://www.cathdb.info/search/by_fasta/results/<task_id>*

Output:

```
query_fasta=<String> original protein sequence used
  for search
cath_version=<string> version of CATH used for scan
signatures_by_id=<Object> regions of the query that
  match CATH
```

The full specification of `signatures_by_id` can be found on the API Web pages.

BASIC PROTOCOL 2

SEARCH CATH WITH A NEW PROTEIN STRUCTURE

Performing a sequence scan of the CATH database is a fast and reliable method of identifying homologous relationships to a new protein (Redfern et al., 2007). If the three-dimensional coordinates for the new protein structure are available, then it is also possible to perform a structural scan of CATH to identify similarities to known proteins. Searching CATH by structure is a computationally expensive procedure, so a structural scan takes much longer to complete than a sequence scan. However, it can be a useful way of examining remote structural similarities in more detail.

Necessary Resources

Hardware

Workstation with network connections

Software

Javascript-enabled Web browser such as Firefox or Internet Explorer

Files

Protein structure file (PDB format)

1. Direct your browser to the main CATH Web site located at *http://www.cathdb.info* (Fig. 1.28.1).

2. On the top menu bar, click "Search," then click "by PDB structure." These pages allow a protein structure to be scanned against representative domains in CATH. The query protein structure must be saved as a file on the local machine in strict PDB format (*http://www.wwpdb.org/documentation/file-format.php*).

3. Click on the "Browse" button to provide a browser dialogue that allows you to locate the query protein structure PDB file on your local machine. Select the required PDB file, then click the "Submit Query" button on the page to start the scan process (Fig. 1.28.5).

4. Select the PDB chain and scan type.

   After submission, the query structure file is examined, and a number of basic checks are made to ensure the data can be correctly understood. A summary of information is then displayed for each of the polypeptide chains present in the file (Fig. 1.28.6). CATH structure scans are made on a per-chain basis, and this list is used to provide the scan options for each chain.

   Performing a sensitive, high-quality structural search against the library of representative domains in CATH is a computationally expensive process that can take anything from a few minutes to 20 min (depending on the server load and the size and type of query structure). If the intention is simply to find CATH Domains or FunFams that best match the query protein (e.g., rather than specifically analyzing the structural alignments) then, where possible, identifying this homology through a sequence search is a much faster alternative.

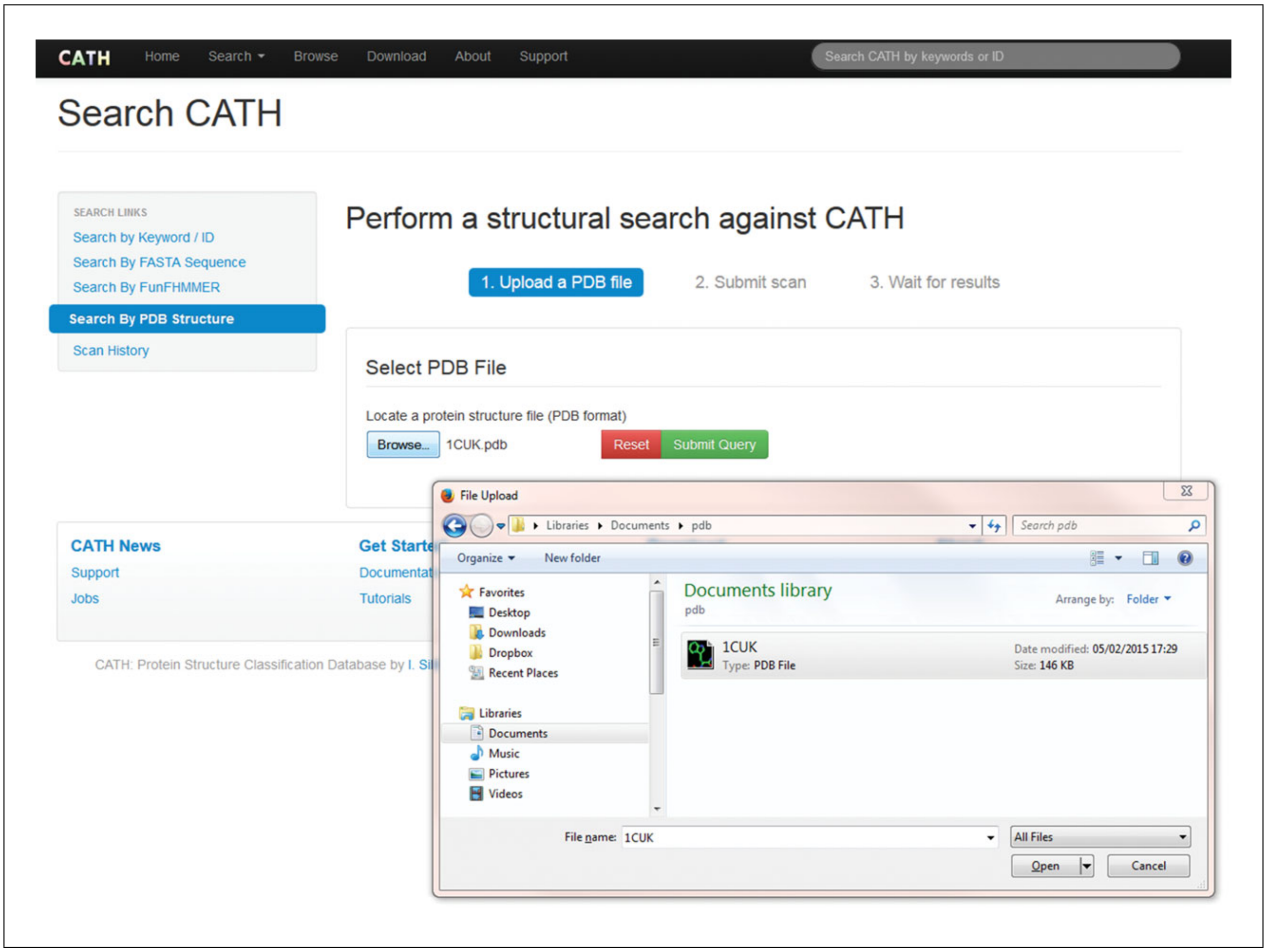


Figure 1.28.5 Submitting a locally stored PDB file to the CATH structural scan.

As a result, the sequence for each chain in the query protein is automatically submitted to the fast CATH sequence scan, and a summary of any available matches appears. Clicking "View sequence hits now" takes the user to a results page as described in "Searching CATH with a new protein sequence." Alternatively, clicking "Submit Structure" submits the query protein chain to a full structural scan. After submitting a structural scan, options are provided either to return to the previous page to submit another scan or move to the "Scan History" page to watch the progress of the scan.

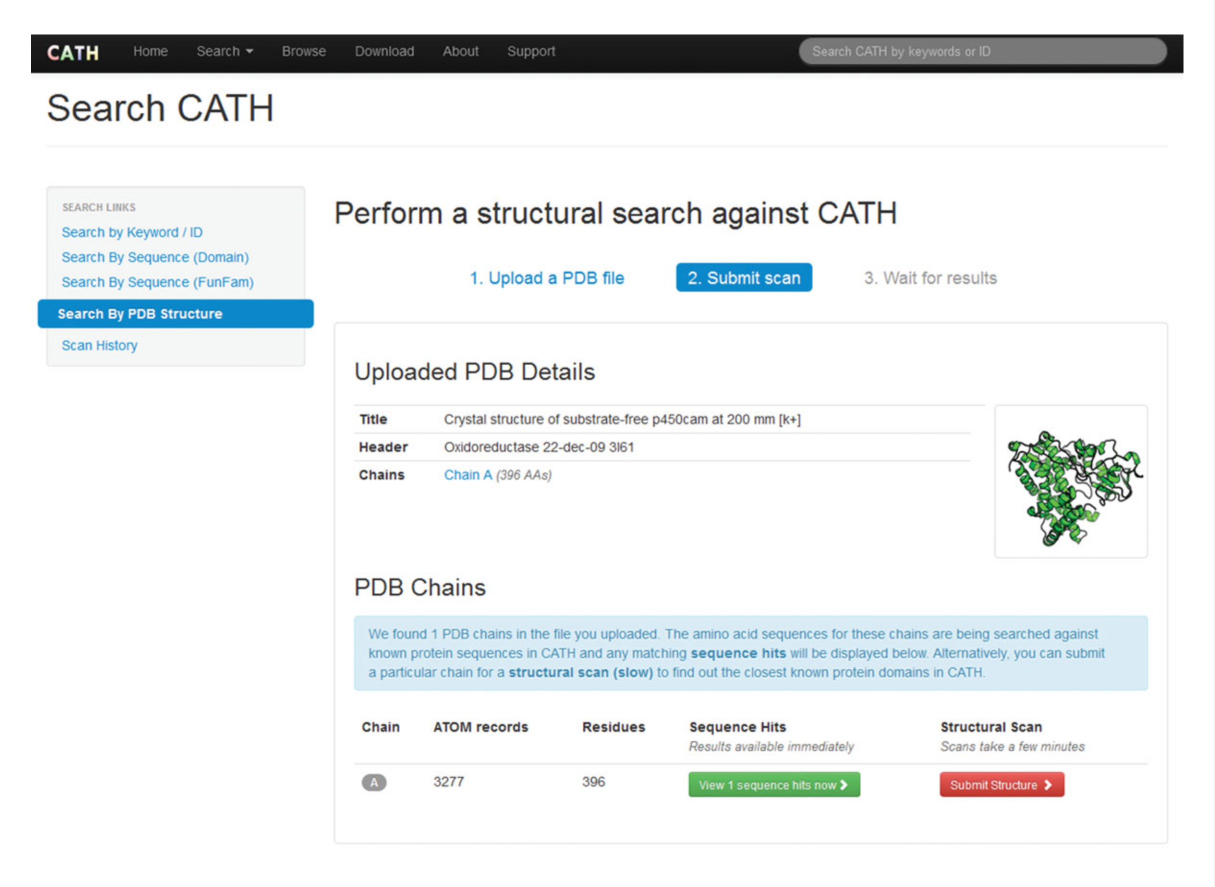
5. View the progress of the structure scan.

The "Scan History" page provides a list of recent scans that have been carried out within this browser session. The status of each scan is provided, and a link to the scan results is displayed as soon as the scan has completed (Fig. 1.28.7).

6. View the results of the structure scan.

The results of matching CATH domains are summarized in a scatter graph (Fig. 1.28.8A). Each colored dot represents a matching CATH domain, and the location is based on two different measures of structural similarity: root-mean-square deviation (RMSD; x-axis) and sequential structure alignment program (SSAP) score (y-axis). The size of the dot is based on the relative percentage of overlap between the query structure and the aligned CATH domain, thus matches that overlap well with the query structure are larger than matches with poor overlap. The color of the dot is calculated by combining the inputs in a sliding, traffic light scale. This provides a visual cue highlighting the most confident matches in green and the least confident matches in red/black.

The details of each CATH domain match are listed below the graph in tabular form (Fig. 1.28.8B). Each entry in this table corresponds to a matching CATH domain with



Search CATH

Perform a structural search against CATH

Uploaded PDB Details

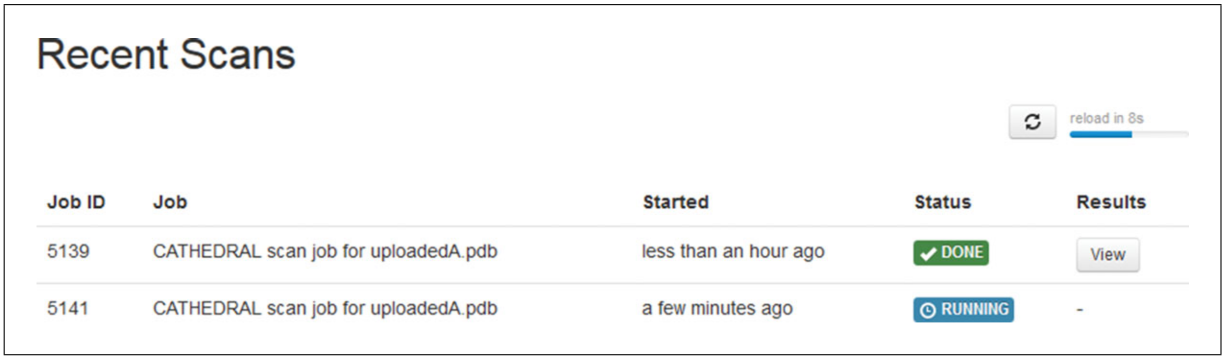
Title	Crystal structure of substrate-free p450cam at 200 mm [k+]
Header	Oxidoreductase 22-dec-09 3l61
Chains	Chain A *(396 AAs)*

PDB Chains

We found 1 PDB chains in the file you uploaded. The amino acid sequences for these chains are being searched against known protein sequences in CATH and any matching **sequence hits** will be displayed below. Alternatively, you can submit a particular chain for a **structural scan (slow)** to find out the closest known protein domains in CATH.

Chain	ATOM records	Residues	Sequence Hits *Results available immediately*	Structural Scan *Scans take a few minutes*
A	3277	396	View 1 sequence hits now	Submit Structure

Figure 1.28.6 Choosing the protein chain and type of scan.



Recent Scans

Job ID	Job	Started	Status	Results
5139	CATHEDRAL scan job for uploadedA.pdb	less than an hour ago	✓ DONE	View
5141	CATHEDRAL scan job for uploadedA.pdb	a few minutes ago	RUNNING	-

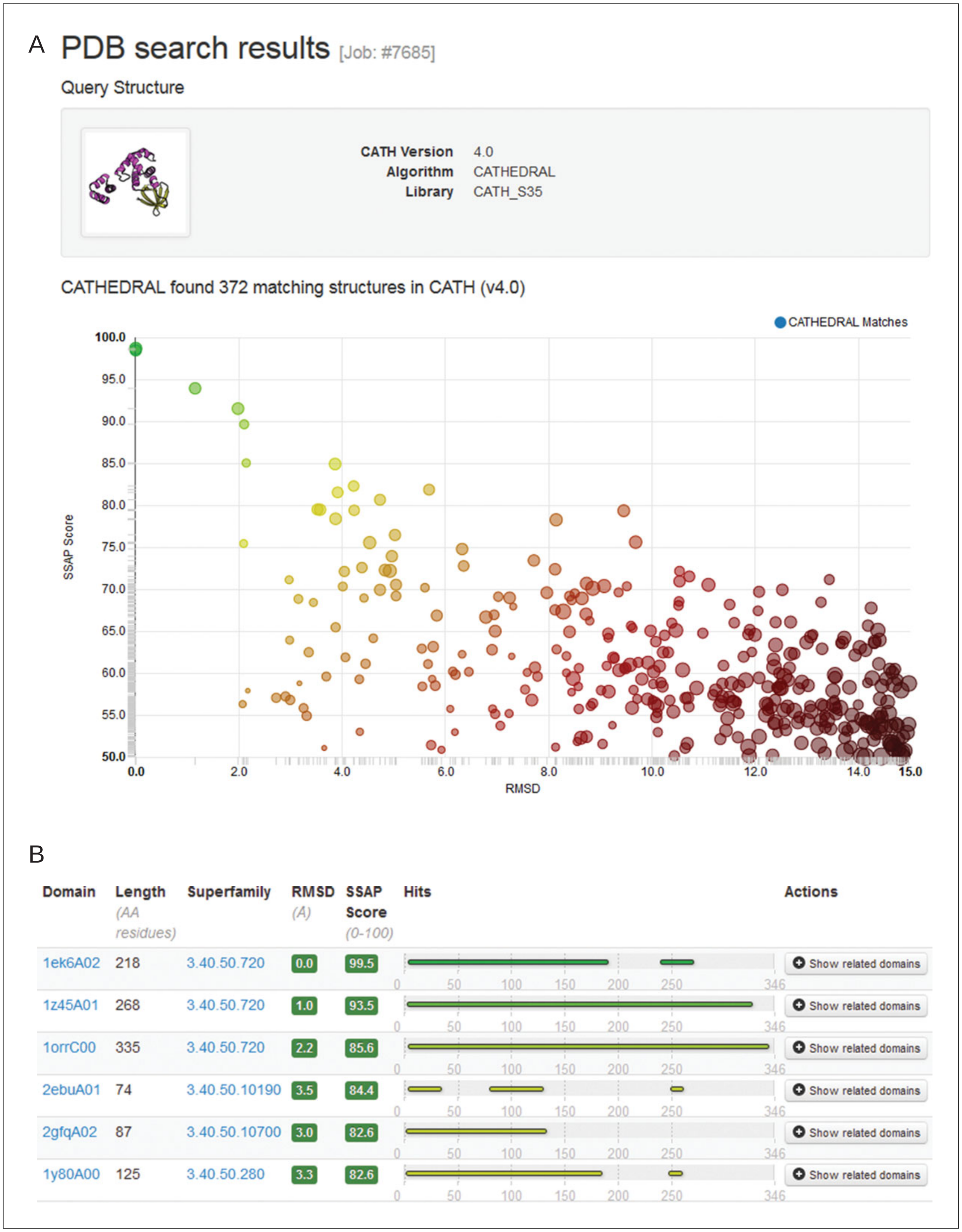
Figure 1.28.7 The Scan History page provides a summary of the progress of recent scans including links to results.

links to the superfamily, the exact RMSD and SSAP score, and an overview of the location of the matching regions of structure. The color of the RMSD and the SSAP score provides an independent assessment of each score, with the combined score giving the final color used for the matching regions and the scatter plot above.

GUIDELINES FOR UNDERSTANDING RESULTS

The main motivation for searching a new protein against CATH is to identify homologous relationships to proteins, or sets of proteins, that have already been well annotated in the database. In CATH, it is possible to identify relationships to a collection of functionally similar domains (functional family or FunFam) or a collection of more distantly related domains (homologous superfamily).

The following sections provide an overview for the information available on the CATH Web pages once relationships to these collections of domains have been identified.



PDB search results [Job: #7685]

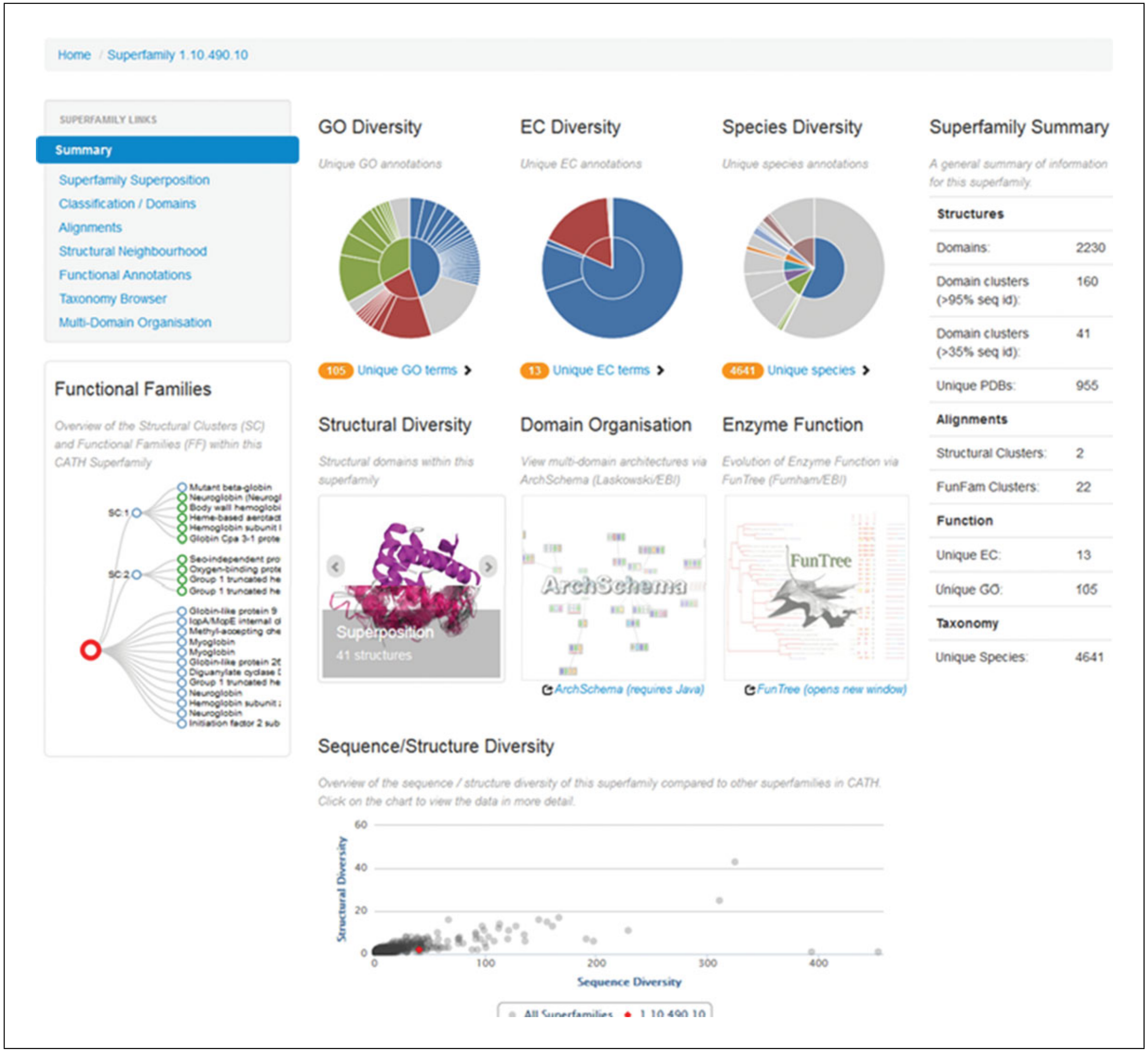
Query Structure

CATH Version	4.0
Algorithm	CATHEDRAL
Library	CATH_S35

CATHEDRAL found 372 matching structures in CATH (v4.0)

Domain	Length (AA residues)	Superfamily	RMSD (Å)	SSAP Score (0-100)	Hits	Actions
1ek6A02	218	3.40.50.720	0.0	99.5		Show related domains
1z45A01	268	3.40.50.720	1.0	93.5		Show related domains
1orrC00	335	3.40.50.720	2.2	85.6		Show related domains
2ebuA01	74	3.40.50.10190	3.5	84.4		Show related domains
2gfqA02	87	3.40.50.10700	3.0	82.6		Show related domains
1y80A00	125	3.40.50.280	3.3	82.6		Show related domains

Figure 1.28.8 (**A**) Results of the structural scan. Each dot represents a matching CATH domain. The size, position, and color provide information on the quality of each match. (**B**) Details of structural scan hits show additional information.



Unique GO annotations

Unique EC annotations

Unique species annotations

Overview of the Structural Clusters (SC) and Functional Families (FF) within this CATH Superfamily

Structural domains within this superfamily

View multi-domain architectures via ArchSchema (Laskowski/EBI)

Evolution of Enzyme Function via FunTree (Furnham/EBI)

Overview of the sequence / structure diversity of this superfamily compared to other superfamilies in CATH. Click on the chart to view the data in more detail.

Superfamily Summary

A general summary of information for this superfamily.

Structures	
Domains:	2230
Domain clusters (>95% seq id):	160
Domain clusters (>35% seq id):	41
Unique PDBs:	955
Alignments	
Structural Clusters:	2
FunFam Clusters:	22
Function	
Unique EC:	13
Unique GO:	105
Taxonomy	
Unique Species:	4641

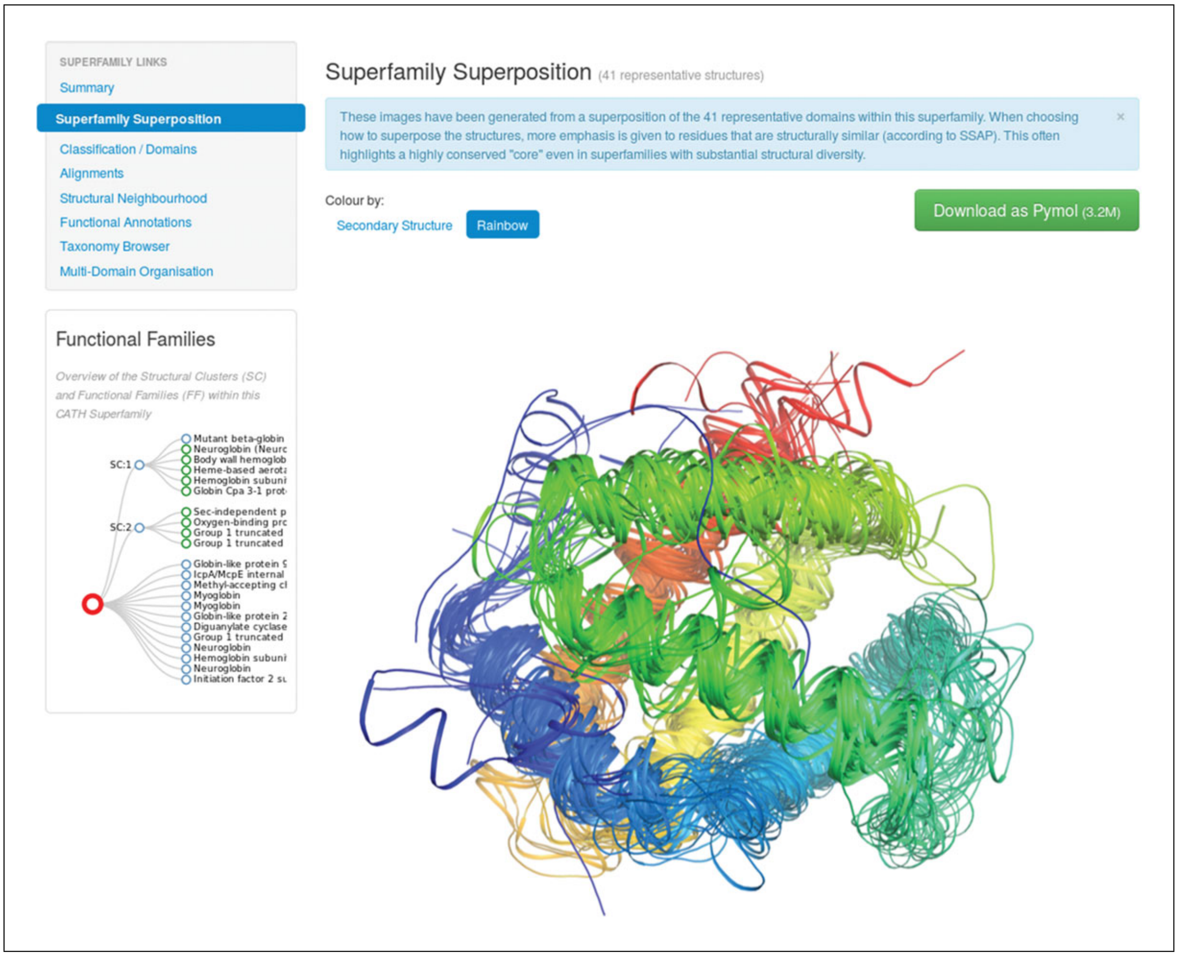
Figure 1.28.9 The CATH superfamily Summary page provides a summary dashboard containing an overview of the information available for this superfamily.

CATH Superfamily Web Pages

The Web pages representing a CATH superfamily provide access to a great deal of information: the classification of structural domains, the conservation and diversity of structure and function, the functional families, domain organizations, and more. The menu on the left hand side of these pages provides access to the different categories of information about this superfamily. Under this menu there is a dendrogram outlining all the FunFams within the superfamily. To indicate which FunFams share more structural similarity than others, the FunFam containing structural CATH domains have been grouped into structural clusters (SCs). Moving the mouse over each node in this dendrogram provides more information, and clicking on the node takes the user to a page dedicated to the individual FunFam.

The list below provides a highlight of the information available from the links on the left hand menu.

1. Superfamily: Summary
   This is the main landing page for the superfamily, containing a broad overview of the features and statistics of the superfamily (Fig. 1.28.9).



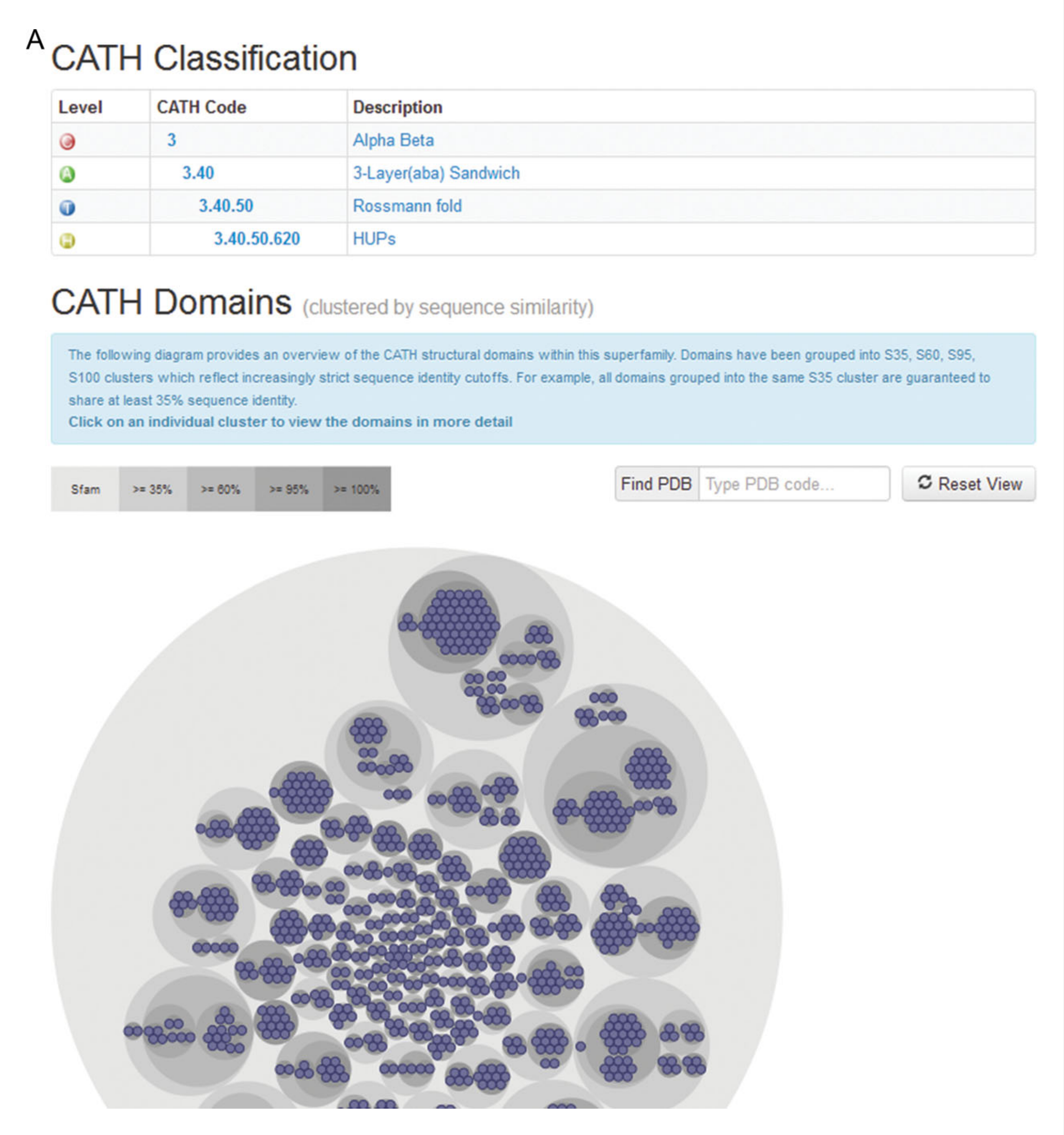
These images have been generated from a superposition of the 41 representative domains within this superfamily. When choosing how to superpose the structures, more emphasis is given to residues that are structurally similar (according to SSAP). This often highlights a highly conserved "core" even in superfamilies with substantial structural diversity.

Figure 1.28.10 The Superfamily Superposition page provides superpositions of all representative domains within a superfamily.

The top row provides a set of pie charts outlining the diversity of functional annotations (in terms of GO terms and EC terms) and taxonomy of all the protein sequences with domains assigned to this CATH superfamily. These charts are hierarchical: the outer circle provides a measure of how well populated individual annotations are within the superfamily, and the inner circle clusters these terms into general categories. Moving the mouse over a particular sector describes more information; the link underneath the chart takes the user to the "Functional Annotation" page listing all these annotations in more detail.

The second row displays images that show a selection of domain structures within this superfamily: the largest, smallest, and representative domain and a superposition of all domains in the superfamily (see Superfamily: Superposition below). The next two features are links to two external resources that provide information on CATH superfamilies: ArchSchema (Tamuri and Laskowski, 2010) which describes the sequence organisation of domains within proteins and FunTree (Furnham et al., 2012) which links the phylogenetic relationships with enzyme classification.

The final row displays a scatter plot indicating the relative sequence and structural diversity of this superfamily (red dot) in comparison to all other superfamilies in CATH (gray dots). Structural diversity is measured as the number of SCs within the



CATH Classification

Level	CATH Code	Description
C	3	Alpha Beta
A	3.40	3-Layer(aba) Sandwich
T	3.40.50	Rossmann fold
H	3.40.50.620	HUPs

CATH Domains (clustered by sequence similarity)

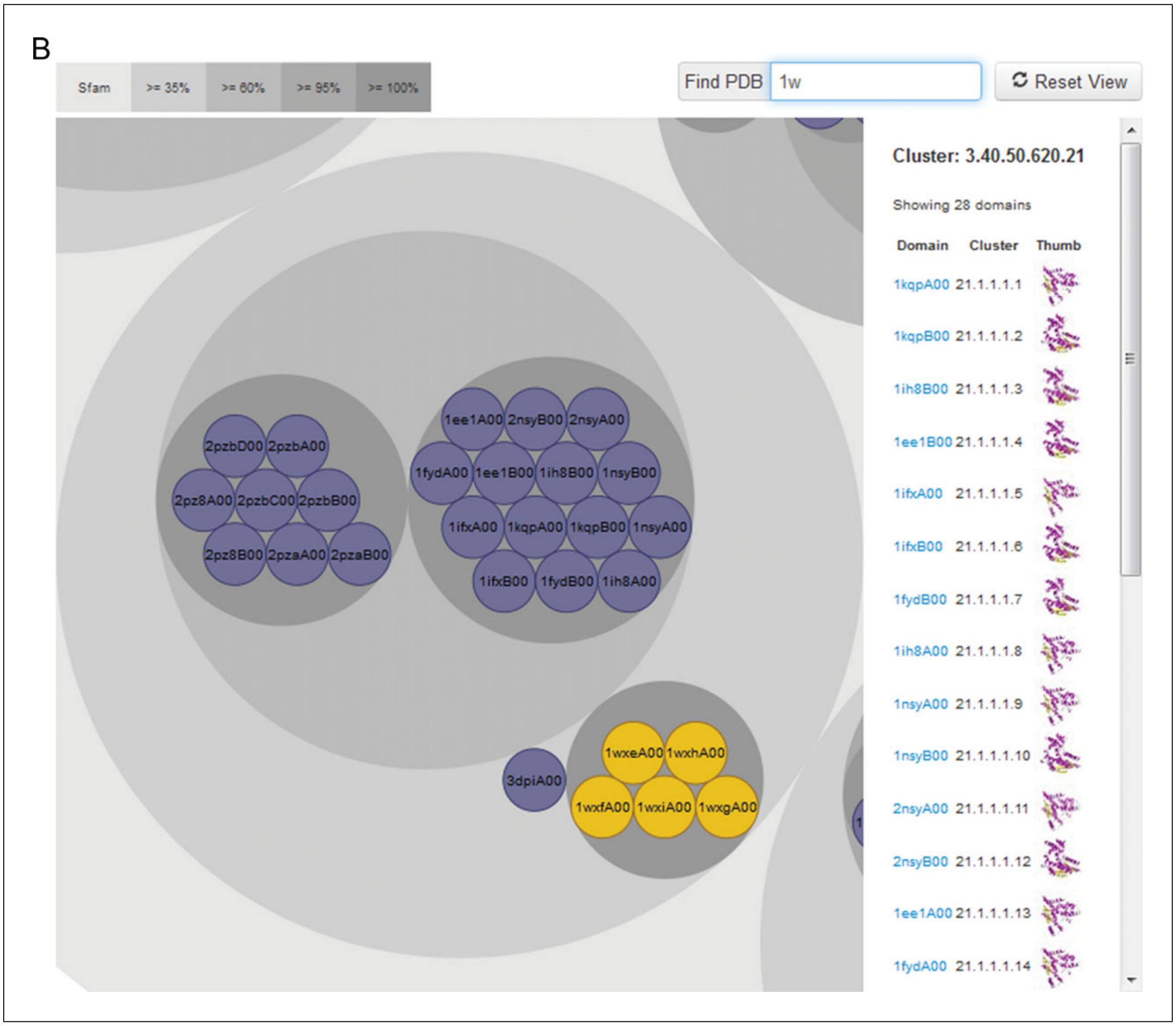
The following diagram provides an overview of the CATH structural domains within this superfamily. Domains have been grouped into S35, S60, S95, S100 clusters which reflect increasingly strict sequence identity cutoffs. For example, all domains grouped into the same S35 cluster are guaranteed to share at least 35% sequence identity.

Click on an individual cluster to view the domains in more detail

Figure 1.28.11 (**A**) The Superfamily Classification page displays the class, architecture, and topology in which the superfamily belongs and the sequence clustering of the member domains. (**B**) Clicking on a gray area zooms into that sequence cluster, while adding text to the "Find PDB" box highlights any matching domain identifications.

superfamily, while sequence diversity is measured as the number of sequence families (i.e., groups of sequences that share ≥35% sequence identity).

2. Superfamily: Superposition
One of the most simple and effective ways of describing a CATH superfamily is to observe the conservation and diversity across its representative domains. To enable this, a structural superposition is provided for each superfamily (Fig. 1.28.10), both as a rendered image (colored by secondary structure and by rainbow) and as a download that can be fed into the PyMOL molecular visualization package (*http://www.pymol.org*).



Cluster: 3.40.50.620.21

Showing 28 domains

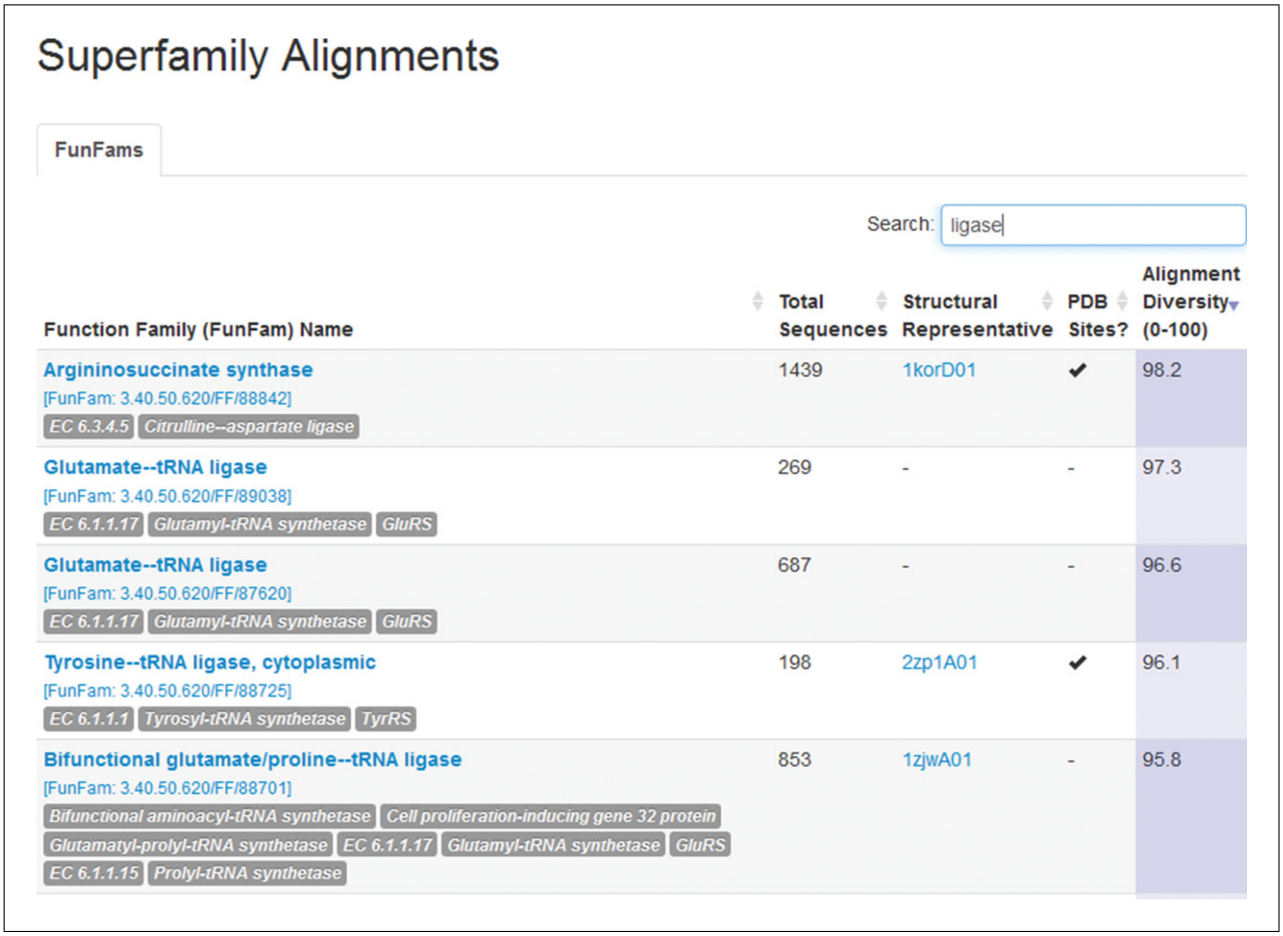
Domain	Cluster	Thumb
1kqpA00	21.1.1.1.1	
1kqpB00	21.1.1.1.2	
1ih8B00	21.1.1.1.3	
1ee1B00	21.1.1.1.4	
1ifxA00	21.1.1.1.5	
1ifxB00	21.1.1.1.6	
1fydB00	21.1.1.1.7	
1ih8A00	21.1.1.1.8	
1nsyA00	21.1.1.1.9	
1nsyB00	21.1.1.1.10	
2nsyA00	21.1.1.1.11	
2nsyB00	21.1.1.1.12	
1ee1A00	21.1.1.1.13	
1fydA00	21.1.1.1.14	

Figure 1.28.11 Continued

3. Superfamily: Classification

   The "Classification" page provides information on the hierarchical lineage (i.e., the class, architecture, and fold to which this superfamily belongs) and an overview of the structural domains that have been assigned to this superfamily. The largest superfamilies in CATH contain many thousands of structural domains, so visualizing and navigating this data in a meaningful way can present a challenge.

   In an effort to organize this data, CATH groups structural domains into sequence families, clustering together domains that share certain levels of sequence identity (35%, 60%, 95%, and 100%). This organization is illustrated as a hierarchical bubble diagram (Fig. 1.28.11A); each domain is represented as a blue dot and the various levels of sequence clustering are represented by the gray circles (moving from light to dark gray as the sequence identity of the member domains increases). Clicking on a cluster (gray circle) will zoom into that cluster and provide a detailed list of member domains. The view can be reset by a button on the top right of the diagram. Typing into the text box marked "Find PDB" highlights any domains matching that text string (changing their color from blue to orange), thus making it possible to quickly locate domains of interest by their PDB identifier (Fig. 1.28.11B).



Superfamily Alignments

Function Family (FunFam) Name	Total Sequences	Structural Representative	PDB Sites?	Alignment Diversity (0-100)
Argininosuccinate synthase [FunFam: 3.40.50.620/FF/88842] EC 6.3.4.5 Citrulline--aspartate ligase	1439	1korD01	✔	98.2
Glutamate--tRNA ligase [FunFam: 3.40.50.620/FF/89038] EC 6.1.1.17 Glutamyl-tRNA synthetase GluRS	269	-	-	97.3
Glutamate--tRNA ligase [FunFam: 3.40.50.620/FF/87620] EC 6.1.1.17 Glutamyl-tRNA synthetase GluRS	687	-	-	96.6
Tyrosine--tRNA ligase, cytoplasmic [FunFam: 3.40.50.620/FF/88725] EC 6.1.1.1 Tyrosyl-tRNA synthetase TyrRS	198	2zp1A01	✔	96.1
Bifunctional glutamate/proline--tRNA ligase [FunFam: 3.40.50.620/FF/88701] Bifunctional aminoacyl-tRNA synthetase Cell proliferation-inducing gene 32 protein Glutamatyl-prolyl-tRNA synthetase EC 6.1.1.17 Glutamyl-tRNA synthetase GluRS EC 6.1.1.15 Prolyl-tRNA synthetase	853	1zjwA01	-	95.8

Figure 1.28.12 The Superfamily Alignments page summaries the FunFams within the superfamily and their alignments.

4. Superfamily: Alignments

The "Alignments" section of the superfamily pages provides a table summarizing all the FunFams within the superfamily (Fig. 1.28.12). The first column contains the FunFam name and keywords (taken from the UniProt description of the most representative sequence). The second column details the number of sequences held in this cluster. If the FunFam happens to contain one or more structural CATH domains, then a representative domain is displayed in the third column. The fourth column declares whether any residue-based annotations are known about active sites in the FunFam. The final column provides an indication of the diversity of the alignment that the FunFam is based on. These scores are calculated using the Scorecons algorithm (Valdar, 2002) and range from 0 (all alignment positions completely identical) to 100 (all alignment positions completely different).

The table can be customized dynamically; the rows can be filtered and sorted without reloading the Web page (along with other tabular content on the CATH site). Entering text into the text box labelled "Search" will filter out any row that does not contain that string. Deleting this search text will bring back all the table data. In addition, clicking on the column headers will sort the table rows based on that column. Clicking again will toggle ascending and descending ordering. Some superfamilies contain a large number of FunFams, so this functionality can be useful to narrow down a list of FunFams of particular interest.

5. Superfamily: Functional Annotations

Details of all the functional annotations for proteins with domains assigned to this superfamily are displayed in the "Functional Annotations" page. These annotations are

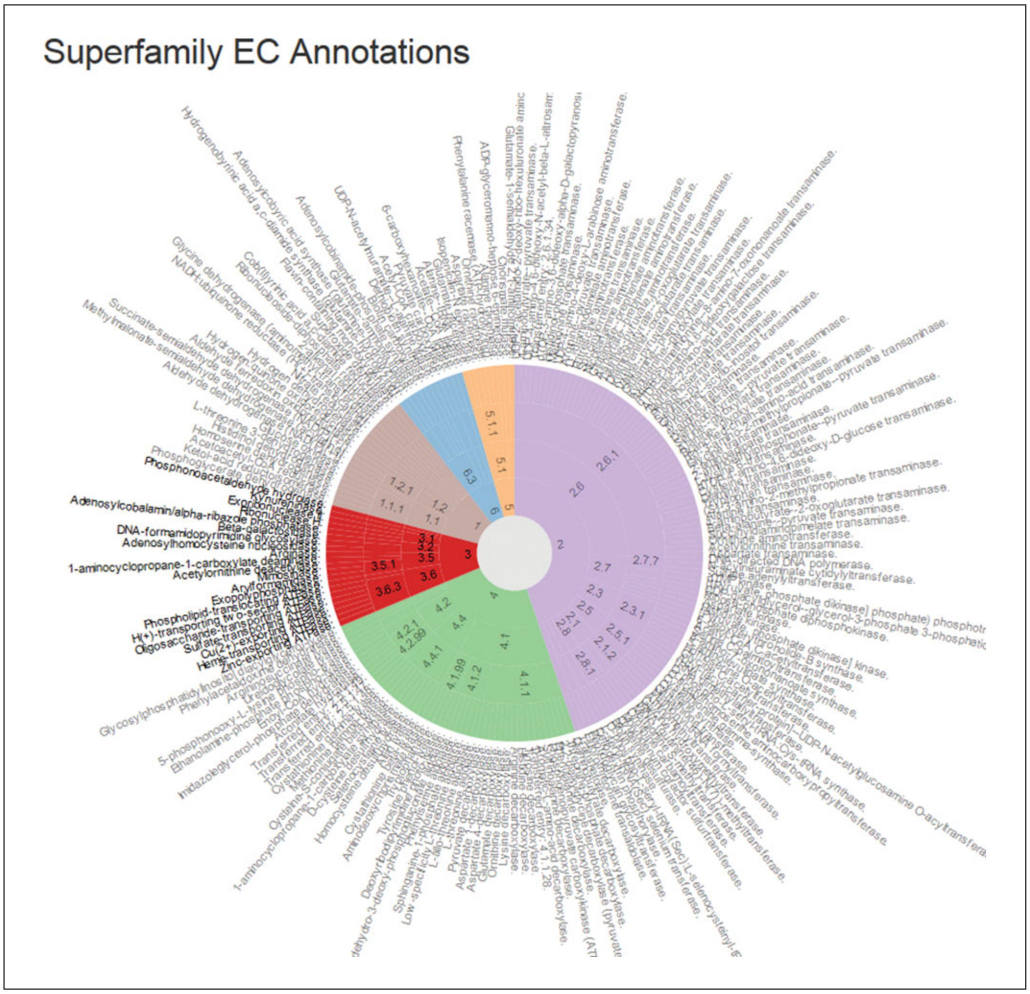
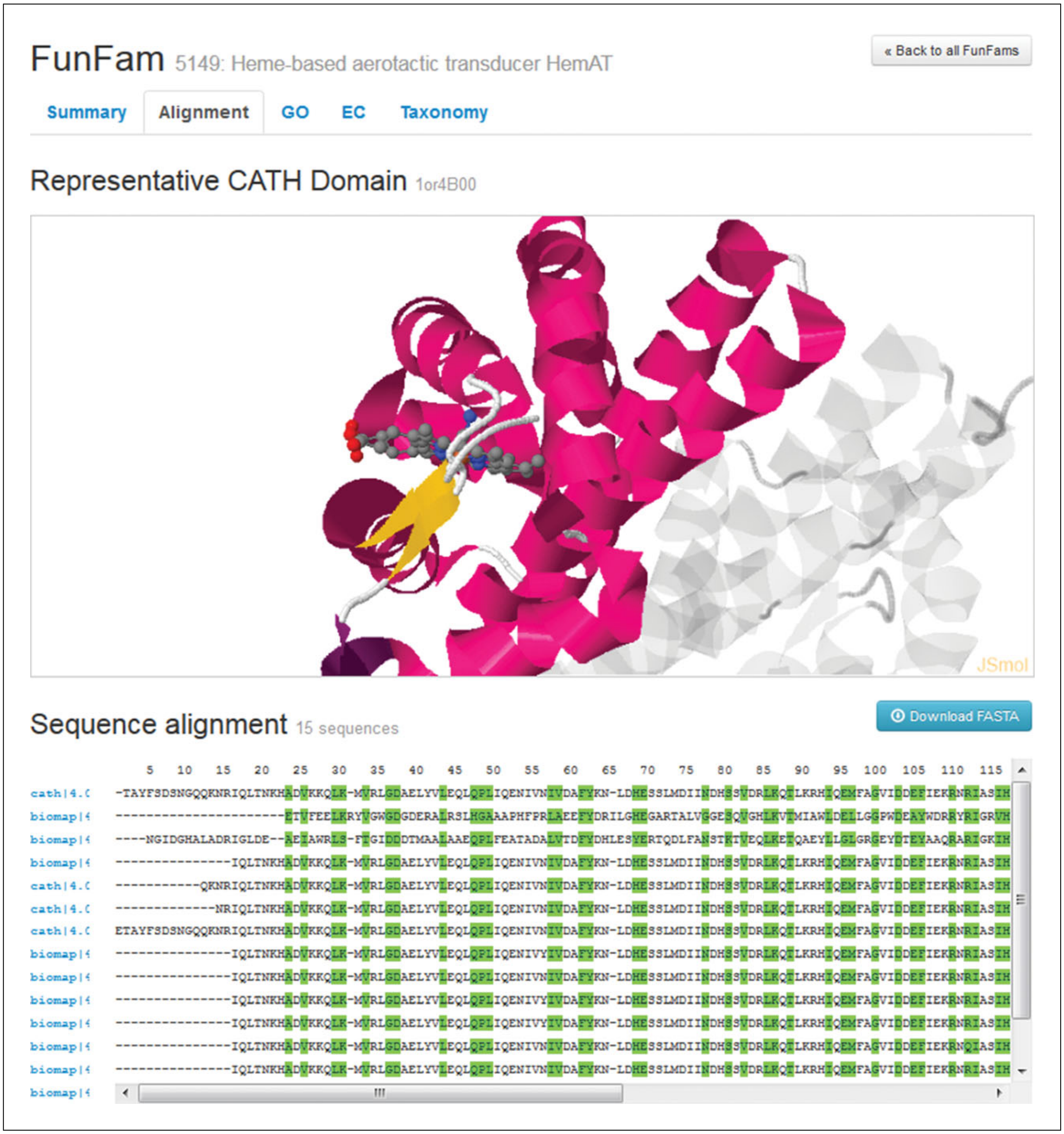


Figure 1.28.13 The Superfamily EC Annotations display EC annotations in a sunburst diagram, with EC classes in the centre and individual descriptions radiating outwards.

split into two tabs according to the source: GO Terms and EC Terms. GO annotations are further separated into broad categories of biological process, cellular components, and molecular function. Clicking the “ReviGO” link submits these GO terms to the ReviGO server (Supek et al., 2011) which provides tools for further visualization and analysis. The EC resource classifies terms describing enzyme functions into a tree-like structure, with highly detailed descriptions of enzyme function at the leaf node that are grouped into more general categories moving towards the trunk (Fig. 1.28.13). This organization is reflected as a sunburst diagram with the six different EC classes in the center of the circle and the individual descriptions in the outer ring. Clicking on any sector will zoom into that area of the chart; clicking on the center of the circle moves back through the tree to the root node. Each of these pages also contains a table (which can be sorted and filtered) with each functional annotation listed in full, along with the number of times this annotation appears in the superfamily.

CATH FunFam Web Pages

The set of Web pages describing CATH functional families share some similarity with the CATH superfamily pages in that they both describe collections of CATH domains. Thus, sections of the dashboard and subsequent detail pages are similar to the superfamily pages: function annotations (GO terms), enzyme annotations (EC terms), and taxonomy.



FunFam 5149: Heme-based aerotactic transducer HemAT

Representative CATH Domain 1or4B00

Sequence alignment 15 sequences

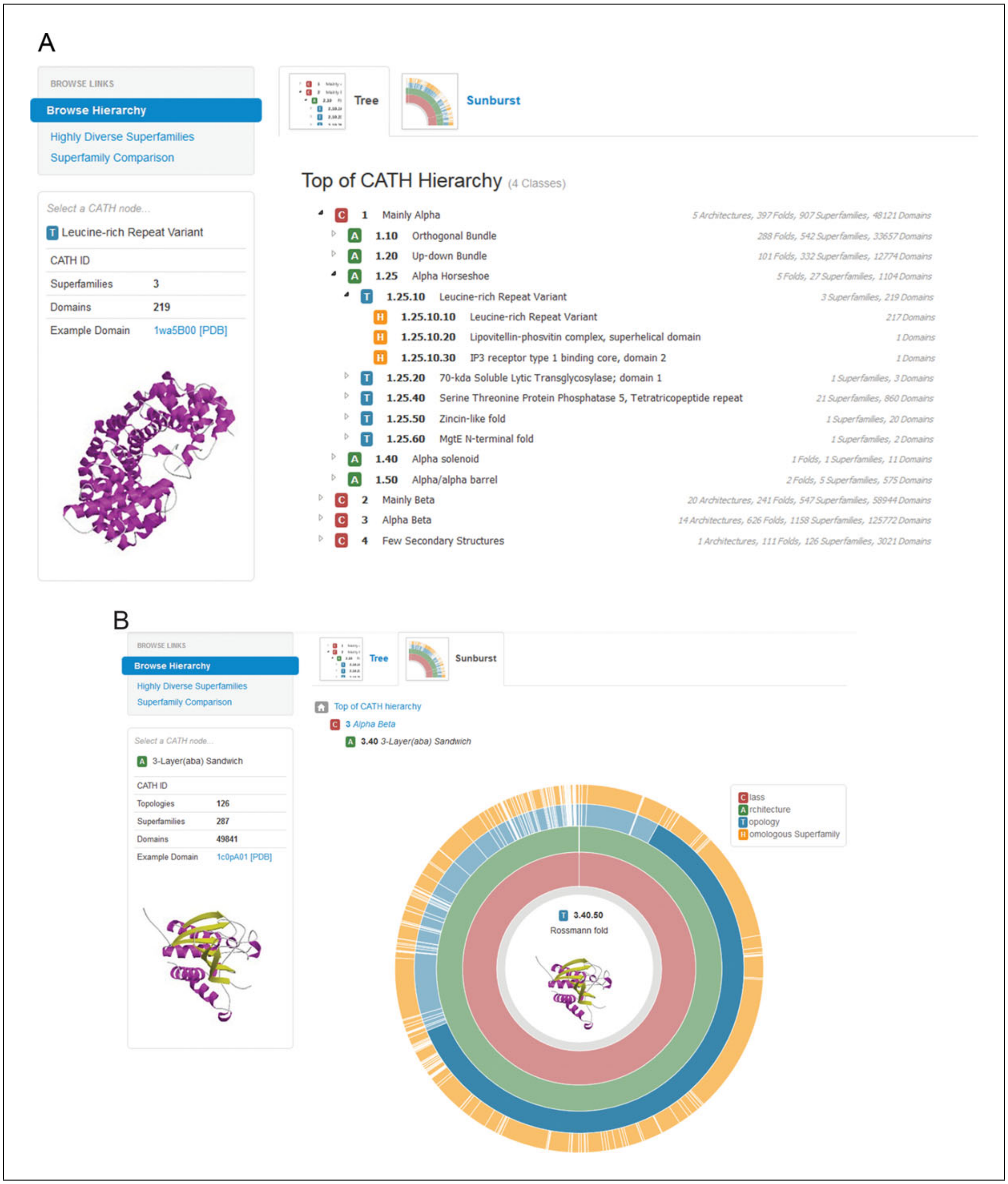
```
              5    10   15   20   25   30   35   40   45   50   55   60   65   70   75   80   85   90   95  100  105  110  115
cath|4.(   -TAYFSDSNGQQKNRIQLTNKHADVKKQLK-MVRLGDAELYVLEQLQPLIQENIVNIVDAFYKN-LDHESSLMDIINDHSSVDRLKQTLKRHIQEMFAGVIDDEFIEKRNRIASIH
biomap|4   ----------------------ETVFEELKRYVGWGDGDERALRSLHGAAAPHFPRLAEEFYDRILGHEGARTALVGGESQVGHLKVTMIAWLDELLGGPWDEAYWDRRYRIGRVH
biomap|4   ----NGIDGHALADRIGLDE--AEIAWRLS-FTGIDDDTMAALAAEQPLFEATADALVTDFYDHLESYERTQDLFANSTKTVEQLKETQAEYLLGLGRGEYDTEYAAQRARIGKIH
biomap|4   ---------------IQLTNKHADVKKQLK-MVRLGDAELYVLEQLQPLIQENIVNIVDAFYKN-LDHESSLMDIINDHSSVDRLKQTLKRHIQEMFAGVIDDEFIEKRNRIASIH
cath|4.(   -----------QKNRIQLTNKHADVKKQLK-MVRLGDAELYVLEQLQPLIQENIVNIVDAFYKN-LDHESSLMDIINDHSSVDRLKQTLKRHIQEMFAGVIDDEFIEKRNRIASIH
cath|4.(   -------------NRIQLTNKHADVKKQLK-MVRLGDAELYVLEQLQPLIQENIVNIVDAFYKN-LDHESSLMDIINDHSSVDRLKQTLKRHIQEMFAGVIDDEFIEKRNRIASIH
cath|4.(   ETAYFSDSNGQQKNRIQLTNKHADVKKQLK-MVRLGDAELYVLEQLQPLIQENIVNIVDAFYKN-LDHESSLMDIINDHSSVDRLKQTLKRHIQEMFAGVIDDEFIEKRNRIASIH
biomap|4   ---------------IQLTNKHADVKKQLK-MVRLGDAELYVLEQLQPLIQENIVYIVDAFYKN-LDHESSLMDIINDHSSVDRLKQTLKRHIQEMFAGVIDDEFIEKRNRIASIH
biomap|4   ---------------IQLTNKHADVKKQLK-MVRLGDAELYVLEQLQPLIQENIVNIVDAFYKN-LDHESSLMDIINDHSSVDRLKQTLKRHIQEMFAGVIDDEFIEKRNRIASIH
biomap|4   ---------------IQLTNKHADVKKQLK-MVRLGDAELYVLEQLQPLIQENIVYIVDAFYKN-LDHESSLMDIINDHSSVDRLKQTLKRHIQEMFAGVIDDEFIEKRNRIASIH
biomap|4   ---------------IQLTNKHADVKKQLK-MVRLGDAELYVLEQLQPLIQENIVYIVDAFYKN-LDHESSLMDIINDHSSVDRLKQTLKRHIQEMFAGVIDDEFIEKRNRIASIH
biomap|4   ---------------IQLTNKHADVKKQLK-MVRLGDAELYVLEQLQPLIQENIVNIVDAFYKN-LDHESSLMDIINDHSSVDRLKQTLKRHIQEMFAGVIDDEFIEKRNQIASIH
biomap|4   ---------------IQLTNKHADVKKQLK-MVRLGDAELYVLEQLQPLIQENIVNIVDAFYKN-LDHESSLMDIINDHSSVDRLKQTLKRHIQEMFAGVIDDEFIEKRNRIASIH
biomap|4
```

Figure 1.28.14 The Alignment tab of the FunFam pages shows the alignment of the FunFam members. Highly conserved positions are shown green and (where possible) can be mapped onto a domain structure.

Additionally, the "Alignment" page (Fig. 1.28.14) highlights the alignment of the members of this functional family with highly conserved positions highlighted in green (calculated with the Scorecons algorithm; Valdar, 2002). If the FunFam contains at least one CATH domain, then these conserved positions can also be visualized on a domain structure that has been selected to represent the FunFam.

Browse the CATH Hierarchy

Within the CATH hierarchy, the homologous superfamily (H-level in CATH) is the most biologically meaningful, as it groups together domains that are related by evolution. However, CATH also attempts to provide an overview of the universe of known protein structural domains by gathering together superfamilies that share different levels of



T Leucine-rich Repeat Variant

CATH ID	
Superfamilies	3
Domains	219
Example Domain	1wa5B00 [PDB]

Top of CATH Hierarchy (4 Classes)

C	1	Mainly Alpha	5 Architectures, 397 Folds, 907 Superfamilies, 48121 Domains
A	1.10	Orthogonal Bundle	288 Folds, 542 Superfamilies, 33657 Domains
A	1.20	Up-down Bundle	101 Folds, 332 Superfamilies, 12774 Domains
A	1.25	Alpha Horseshoe	5 Folds, 27 Superfamilies, 1104 Domains
T	1.25.10	Leucine-rich Repeat Variant	3 Superfamilies, 219 Domains
H	1.25.10.10	Leucine-rich Repeat Variant	217 Domains
H	1.25.10.20	Lipovitellin-phosvitin complex, superhelical domain	1 Domains
H	1.25.10.30	IP3 receptor type 1 binding core, domain 2	1 Domains
T	1.25.20	70-kda Soluble Lytic Transglycosylase; domain 1	1 Superfamilies, 3 Domains
T	1.25.40	Serine Threonine Protein Phosphatase 5, Tetratricopeptide repeat	21 Superfamilies, 860 Domains
T	1.25.50	Zincin-like fold	1 Superfamilies, 20 Domains
T	1.25.60	MgtE N-terminal fold	1 Superfamilies, 2 Domains
A	1.40	Alpha solenoid	1 Folds, 1 Superfamilies, 11 Domains
A	1.50	Alpha/alpha barrel	2 Folds, 5 Superfamilies, 575 Domains
C	2	Mainly Beta	20 Architectures, 241 Folds, 547 Superfamilies, 58944 Domains
C	3	Alpha Beta	14 Architectures, 626 Folds, 1158 Superfamilies, 125772 Domains
C	4	Few Secondary Structures	1 Architectures, 111 Folds, 126 Superfamilies, 3021 Domains

A 3-Layer(aba) Sandwich

CATH ID	
Topologies	126
Superfamilies	287
Domains	49841
Example Domain	1c0pA01 [PDB]

Top of CATH hierarchy
C 3 *Alpha Beta*
A 3.40 *3-Layer(aba) Sandwich*

Figure 1.28.15 (**A**) Viewing the CATH hierarchy with the tree browser. (**B**) Browsing the CATH hierarchy with the sunburst browser. This diagram shows all nodes in the CATH hierarchy with the four main classes in the inner circle (red), then architectures (green), topologies (blue), and homologous superfamilies (yellow) in the outer circle.

similarity. This hierarchy can be browsed on the Web page in two different formats: *tree* or *sunburst*.

1. Viewing the CATH hierarchy with the tree browser
   The tree browser allows the user to browse through the hierarchy as if it were a directory structure of a file system (Fig. 1.28.15A). At the "trunk" of the tree are the four main CATH classes (C-level). Each class is split into architectures (A-level), which are split into topologies (T-level), and finally homologous superfamilies

(H-level). Clicking on an entry will open/close the listing of that node and display more information in the window on the left hand side of the screen.

2. Viewing the CATH hierarchy with the sunburst diagram
   An alternative method of visualising this CATH hierarchy is through a sunburst diagram (Fig. 1.28.15B). In this case, the root of the tree is represented by the innermost circle of the diagram (in gray). Moving towards the outside of the circle, the root circle is then split into the four main CATH classes (in red) which are each split into architectures (green), topologies (blue), and homologous superfamilies (yellow). By default, the relative size of each node in the hierarchy is governed by the number of domains contained within that branch of the tree.

   Moving the mouse over each node adds a preview of the representative domain structure in the centre of the circle. Clicking on a particular node reconfigures the diagram to only show nodes within that branch of the tree. Thus, clicking the architecture "3-layer (aba) Sandwich" (CATH architecture 3.40) will reconfigure the diagram to only show the topologies and homologous superfamilies within that architecture. Clicking on the innermost gray circle will return the diagram back to showing the entire tree.

COMMENTARY

Background Information

The CATH-Gene3D database (Lees et al., 2012; Sillitoe et al., 2013) contains a great deal of data from a number of different sources. In making sense of this information, it is useful to keep in mind the core units of data and how they are generated:

CATH Domain: compact unit of protein structure

CATH Superfamily: collection of domains that are related by evolution

CATH FunFams: collection of domains within a superfamily that have the same or highly similar functions

Where possible, CATH uses an automated pipeline to identify how best to chop PDB structures into domains and how best to assign those domains into related superfamilies. A number of different algorithms are used to provide evidence for this classification process from evaluating homology to known domains (by comparing sequence and structural information) to the ab initio prediction of domain boundaries and comparison of known functional annotations. A set of strictly benchmarked criteria is used to identify cases where this evidence is sufficient to make an assignment without being checked by a human, otherwise the individual chopping and assignments are passed on for manual curation. In CATH version 4.0, 96% of assignments were made automatically leaving 4% to be manually checked.

At the time of writing, we are preparing version 4.1 of CATH which is expected to have around 2,750 superfamilies and over 300,000 structural domains including over 30 million protein sequences.

Suggestions for Further Analysis

To use CATH data programmatically, it is recommended to use the many different types of flat files that CATH provides for download at *http://release.cathdb.info*. Some of the key files are described below; others can be found by exploring within the directory structure.

CATH-B (putative domain assignments between releases)

We have recently started providing access to the putative domain assignments made between CATH releases at *http://release.cathdb.info/cath1_b*. Though users should be aware that these annotations have not been subjected to the validation associated with CATH releases, we expect changes to be rare and to mostly arise from merges reflecting newly identified evolutionary relationships.

The file layout is described in a `README.txt` file. Each date-stamped snapshot contains three files: one for domains that were present in the previous release, one for domains that have been assigned to CATH since then, and one for all. The newest snapshot is also available with a filename that has '`newest`' in place of the date-stamp. Each file (once decompressed with gunzip or similar) contains one line per domain of the format:

```
  domain_id status putative_
superfamily_id putative_
chopping
```

CATH release files

All other downloadable flat files are associated with CATH releases, so the first step to access them is to choose a release of CATH and click the corresponding subdirectory (or 'latest_release'). Highlights include:

1. A strictly non-redundant set of domains, identified using BLAST at 40% sequence identity and 60% overlap. This is provided as: sequences, a list of domain identifications, and a tar.gz of PDB files. The relevant files (including an explanatory README.txt) have names beginning Cath.DataSet.NonRedundant.S40_overlap_60.
2. README.*: a set of files describing the details of the file formats
3. CathDomainList: a one-domain-per-line summary of all assigned CATH domains and their CATH classifications
4. CathDomainDescriptionFile: a detailed description of each assigned CATH domain
5. CathDomall: a description of the CATH domain boundaries for each PDB chain, specified in terms of PDB residue numbers
6. CathDomall.seqreschopping: a description of the CATH domain boundaries for each PDB chain, specified in terms of the SEQRES residue numbers
7. CathNames: a one-node-per-line summary of all CATH nodes (i.e., classes, architectures, topologies, and homologous superfamilies) and their names
8. SequenceBySuperfamily: a directory of FASTA files of sequences for each superfamily (e.g., CathSuperfamilySeqs.1.10.8.70.COMBS.v4.0.0)

Acknowledgements

This work was supported by the Biotechnology and Biological Sciences Research Council [BB/K020013/1] and the Wellcome Trust [081989/Z/07/Z]. A huge number of people have contributed to the CATH resource over the last twenty years. We would like to acknowledge the efforts of the CATH team (past and present), the scientific advisory board for their guidance and advice, and the users who have contributed through comments, suggestions, and bug reports.

Literature Cited

Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, E.E. Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., and Tasumi, M. 1977. The protein data bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol.* 112:535-542.

Furnham, N., Sillitoe, I., Holliday, G.L., Cuff, A.L., Rahman, S.A., Laskowski, R.A., Orengo, C.A., and Thornton, J.M. 2012. FunTree: A resource for exploring the functional evolution of structurally defined enzyme superfamilies. *Nucl. Acids Res.* 40:D776-D782.

Lees, J., Yeats, C., Perkins, J., Sillitoe, I., Rentzsch, R., Dessailly, B.H., and Orengo C. 2012. Gene3D: A domain-based resource for comparative genomics, functional annotation and protein network analysis. *Nucl. Acids Res.* 40:D465-D471.

Redfern, O.C., Harrison, A., Dallman, T., Pearl, F.M., and Orengo, C.A. 2007. CATHEDRAL: A fast and effective algorithm to predict folds and domain boundaries from multidomain protein structures. *PLoS Comput. Biol.* 3(11):e232.

Sillitoe, I., Cuff, A.L., Dessailly, B.H., Dawson, N.L., Furnham, N., Lee, D., Lees, J.G., Lewis, T.E., Studer, R.A., Rentzsch, R., Yeats, C., Thornton, J.M., and Orengo, C.A. 2013. New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D Structures. *Nucl. Acids Res.*:D490-D498.

Sillitoe, I., Lewis, T.E., Cuff, A., Das, S., Ashford, P., Dawson, N.L., Furnham, N., Laskowski, R.A., Lee, D., Lees, J.G., Lehtinen, S., Studer, R.A., Thornton, J., and Orengo, C.A. 2015. CATH: Comprehensive structural and functional annotations for genome sequences. *Nucl. Acids Res.* 43:D376-D381.

Supek, F., Bošnjak, M., Škunca, N., and Šmuc T. 2011. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One.* 6(7):e21800.

Tamuri, A.U. and Laskowski, R.A. 2010. ArchSchema: A tool for interactive graphing of related Pfam domain architectures. *Bioinformatics* 26:1260-1261.

Valdar, W.S. 2002. Scoring residue conservation. *Proteins* 48:227-241.

Yeats, C., Redfern, O.C., and Orengo C. 2010. A fast and automated solution for accurately resolving protein domain architectures. *Bioinformatics* 26:745-751.