Software review

Exploring protein domain structure

Abstract
The protein databank contains coordinates of over 10,000 protein structures, which constitute more than 25,000 structural domains in total. The investigation of protein structural, functional and evolutionary relationships is fundamental to many important fields in bioinformatics research, and will be crucial in determining the function of the human and other genomes. This review describes the SCOP and CATH databases of protein structure classification, which define, classify and annotate each domain in the protein databank. The hierarchical structure, use and annotation of the databases are explained. Other tools for exploring protein structure relationships are also described.

INTRODUCTION
The coordinates of over 10,000 protein structures have been determined by X-ray crystallography or nuclear magnetic resonance spectroscopy (NMR) and are available in the protein databank, PDB. These data underpin the functional genomics effort to characterise the function of the human and other genomes. They provide template and benchmark datasets for use in fold recognition experiments that attempt to determine which of the known protein structures are compatible with a given novel sequence. In cases where two related proteins have evolved beyond any recognisable sequence similarity, the structural data provide a basis on which to relate the proteins and to infer functional properties. However, comparison of similar protein structures can provide valuable biological insights whether through inference of a distance homologous relationship or by analogy only. Where homology is implied, the known coordinate data provide a template upon which a model structure can be constructed by comparative modelling techniques. Over 200 new protein structures are being determined each month and this rate is likely to increase along with progress in various structural genomics projects. It is fortunate, therefore, that powerful tools exist that both classify and provide a means to explore this ever-growing microcosm of protein folds. The 10,650 entries in the February 2000 release of PDB contained 24,186 structural domains. Domains extracted from PDB are defined, classified and annotated in the SCOP and CATH databases. The databases employ hierarchic classifications generated semi-automatically on the basis of similarity in domain sequence, topology, fold and function. However, the high quality of the data arises from the great effort made by the experts who manually inspect the protein relationships and ensure that a meaningful classification is given for each domain. This review will describe SCOP, CATH and other resources that are available on the WWW for exploring and relating protein domain structures.

DESCRIPTION OF THE CLASSIFICATION HIERARCHIES
SCOP and CATH employ related classification hierarchies (Table 1), in which domains are organised into...
Table 1: Summary of classification hierarchies used in SCOP and CATH

<table>
<thead>
<tr>
<th>SCOP</th>
<th>CATH</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Class</strong> – proteins with similar secondary structure element composition</td>
<td><strong>Class</strong> – proteins with similar secondary structure element composition (segregation of elements is not considered)</td>
</tr>
<tr>
<td><strong>Fold</strong> – proteins with a similar core of secondary structure, arranged in a similar way in space and with the same topology</td>
<td><strong>Architecture</strong> – proteins with a similar shape of fold or the same general packing arrangement of secondary structure elements</td>
</tr>
<tr>
<td><strong>Superfamily</strong> – proteins with a possible common evolutionary origin (proteins share structural and or functional similarities)</td>
<td><strong>Topology</strong> – proteins must share the same connectivity between their secondary structure elements</td>
</tr>
<tr>
<td><strong>Family</strong> – proteins that are clearly homologous</td>
<td><strong>Homologous superfamily</strong> – proteins are likely to be homologous (must possess similar functions and have a threshold level of objective structural similarity)</td>
</tr>
<tr>
<td><strong>Family</strong> – proteins that are clearly homologous</td>
<td><strong>Family</strong> – proteins that are clearly homologous (must share at least 20 per cent residue identity)</td>
</tr>
</tbody>
</table>

The text is a summary of the constraints that have to be satisfied in order for two domains to be grouped together at the appropriate level. Note that the classification for both databases is partly subjective at most levels in the hierarchies.

Tree-like structures of several levels. The first and highest level in SCOP is the protein class. Each class comprises a number of unique folds (the second level), each of which lists a number of superfamilies (level three). Each of these comprises a number of distinct protein families (the lowest level in the tree). The 'leaves' of the tree are the individual protein domains themselves. The levels in CATH are class, architecture, topology, homologous superfamily and sequence family. It is important to consider that the nomenclature mentioned here lacks precise generic meaning and operational definitions, when given, are often subjective. Descriptions of the levels in the two databases now follow: Individual domains are clustered into families if there is a strong suggestion of an evolutionary relationship. In the case of SCOP, two domains must possess at least 30 per cent residue identity or otherwise have a very similar function and structure. For CATH, membership is qualified if at least 35 per cent residue identity is seen or if the proteins have high structural similarity and at least 20 per cent residue identity. Even in the absence of any detectable sequence similarity, domains from two different families may share structural or functional similarities that are suggestive of a common evolutionary origin and indicate that the families may be part of a larger superfamily. In the case of CATH, two domains must possess a threshold level of objective structural similarity, and also possess similar functions, to be placed in the same homologous superfamily. In contrast, the SCOP superfamilies are assigned on a more subjective basis. At the next level (topology) in the CATH hierarchy, domain topology is resolved. Members of two superfamilies must possess the same connectivity between their secondary structure elements to be grouped together at this level. Protein topology is considered at the level of fold in SCOP. Two domains are defined as having the same fold if they possess a similar core of secondary structure elements, which adopt a similar arrangement in space and form the same topological connections. In contrast, two domains do not have to possess the same core secondary
structure elements to possess the same architecture in CATH; it is sufficient if they share a similar shape of fold (eg "barrels") or show the same general packing arrangement of secondary structure elements (eg "helical bundle").

As a consequence, some of the CATH architectures are rather broad and may encompass several SCOP folds. Where similarities are seen at the level of fold or architecture only an evolutionary relationship is not implied. The similarity may have arisen through a process of convergent evolution because certain folding pathways, topologies or packing arrangements are naturally favoured. In SCOP, each fold belongs to a certain class depending on its secondary structure element composition as follows: (i) all-$\alpha$ (structure is mostly $\alpha$-helices), (ii) all-$\beta$ (structure is mostly $\beta$-sheet), (iii) $\alpha$-$\beta$ (mixed composition of $\alpha$-helices and $\beta$-sheet) and (iv) $\alpha+\beta$ ($\alpha$-helices and $\beta$-sheet are mostly segregated in the fold). Additional classes exist for (v) multi-domain proteins (folds consisting of two or more domains belonging to different classes), (vi) membrane and cell surface proteins and peptides, (vii) small proteins, (viii) coiled coil proteins, (ix) low-resolution protein structures, (x) peptides and (xi) designed proteins.

The CATH class level does not consider segregation of secondary structure elements. Three main classes are included (mainly- $\alpha$, mainly- $\beta$ and $\alpha$-$\beta$) with the fourth class of "few secondary structures' reserved for proteins with sparse secondary structure composition.

**NAVIGATING THE HIERARCHIES**

The databases are available as sets of linked hypertext documents on the WWW (see Table 2). Their organisation makes it possible to rapidly identify structural neighbours (proteins with a similar structure) to a domain of interest. After entering either database

<table>
<thead>
<tr>
<th>Table 2: URLs of resources</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Resource</strong></td>
</tr>
<tr>
<td>3DB Browser</td>
</tr>
<tr>
<td>3dSearch</td>
</tr>
<tr>
<td>CE</td>
</tr>
<tr>
<td>Dis</td>
</tr>
<tr>
<td>LIGPLOT</td>
</tr>
</tbody>
</table>
at the top of the hierarchy, the user is confronted with a list of links for the available classes. By selecting a link, the user descends to the next level down in the hierarchy, in the case of SCOP, a list of links for each fold would appear. Each level is linked to the next one down in this manner and a lineage, detailing the current position in the hierarchy, provides a convenient means to navigate back up the tree.

At the bottom level in each tree, the protein domains themselves are listed. In the case of SCOP, domains are organised by domain type and species for each family. An extract of the SCOP classification for the fatty-acid-binding protein-like family is shown in Figure 1. The family comprises ten distinct types of domain. There are examples of the muscle fatty-acid-binding protein from *Homo sapiens* (four examples) and *Bos taurus* (one example). Selecting the link to *Homo sapiens* would bring up a list of links for four PDB files, annotated with descriptions of the proteins state of ligation. Selecting one of these final links would start the PDB Entry Viewer tool (see below) containing annotation for the appropriate domain. For CATH, the PDB code of each domain in the family is provided as links. Selecting a link would start the PDBsum tool (see below) containing annotation for that domain. Both SCOP and CATH also provide search tools for querying the databases by PDB identifier code or keywords.

**DOMAIN ANNOTATION IN SCOP AND CATH**

In addition to classifying sequence, structural and functional relationships, SCOP and CATH provide extremely rich annotation for each domain, mostly in the form of links to other WWW-based sources of relevant information. Most of this annotation is provided by the PDB Entry Viewer (SCOP) and PDBsum (CATH) tools (summarised in Table 3), which include diverse information such as atomic coordinate data, molecular images, secondary structure and many links to external resources. Several other resources for interrogating a PDB file are available, including the PDBFields query tool provided at the PDB website and the OCA Viewer. When used in conjunction with SCOP or CATH, these resources allow detailed protein information to be rapidly acquired for the domain of interest and its structural neighbours. A further level of annotation is provided in CATH by their dictionary of homologous superfamilies, DHS. CATH also includes a lexicon, providing textual descriptions of the top two levels (class and architecture) in the hierarchy. The lexicon is tightly coupled to a useful glossary of terms and list of references.

**FINDING HOMOLOGUES TO KNOWN PROTEIN STRUCTURES**

A common task is to establish whether a novel protein sequence is homologous to a protein of known structure. SCOP provides a fast and convenient means to do this through its use of the intermediate sequence library, PDB-ISL, which is a collection of sequences homologous to known structural domains. PDB-ISL is generated by searching a non-redundant protein sequence database using the PSI-BLAST program with the sequences of known domains as the query sequences. A WWW-based interface for screening PDB-ISL is available (see Table 2). A match to a sequence in the library implies a match to a protein of known structure. A similar library for CATH called the protein family database, CATH-PFDB, is being generated and...
**Family: Fatty acid binding protein-like**

ten-stranded meander beta-sheet folded upon itself
relates to the common fold by opening the barrel and insertion of
beta-hairpin

**Lineage:**
1. Root: scop
2. Class: All beta proteins
3. Fold: Lipocalins
   barrel, closed or opened; n=8, S=12; meander
4. Superfamily: Lipocalins
   bind hydrophobic ligands in their interior
5. Family: Fatty acid binding protein-like
   ten-stranded meander beta-sheet folded upon itself
   relates to the common fold by opening the barrel and insertion of beta-hairpin

**Protein Domains:**
1. Muscle fatty acid binding protein (m-fabp)
   1. Human (*Homo sapiens*) (4)
   2. Bovine (*Bos taurus*) (1)
2. Intestinal fatty acid binding protein
   1. Rat (*Rattus rattus*) (8)
   2. Human (*Homo sapiens*) (1)
3. Epidermal fatty acid binding protein
   1. Human (*Homo sapiens*) (1)
4. Adipocyte lipid-binding protein, ALBP
   1. Mouse (*Mus musculus*) (12)
5. Fatty acid-binding protein
   1. Tobacco hornworm (*Manduca sexta*) (1)
   2. Desert locust (*Schistocerca gregaria*) (1)
6. Cellular retinoid-binding protein (CRABP)
   1. Human (*Homo sapiens*), CRABP-II (7)
   2. Bovine/murine (identical sequences) (*Bos taurus/Mus musculus*), CRABP-1 (3)
7. Cellular retinoil-binding protein II (CRBP)
   1. Rat (*Rattus rattus*) (4)
8. Liver fatty acid binding protein
   1. Rat (*Rattus norvegicus*) (1)
9. P2 myelin protein
   1. Bovine (*Bos taurus*), caudal spinal root myelin (1)
10. Ileal lipid binding protein
    1. Pig (*Sus scrofa*) (1)

Figure 1: Extract from the SCOP classification for the fatty-acid-binding protein-like family. The family adopts the lipocalin fold, which is mostly beta-sheet structure. The family comprises ten distinct types of domain (muscle fatty-acid-binding protein, intestinal fatty-acid-binding protein, etc.). PDB files for four muscle fatty-acid-binding proteins from *Homo sapiens* and one from *Bos taurus* are available.
Table 3: Annotation provided by SCOP and CATH

<table>
<thead>
<tr>
<th>Annotation</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SCOP (PDB Entry Viewer)</strong></td>
<td></td>
</tr>
<tr>
<td>SCOP Classification</td>
<td>Link to entry in the SCOP database</td>
</tr>
<tr>
<td>Header</td>
<td>Displays header information from PDB file</td>
</tr>
<tr>
<td>Full Text</td>
<td>Displays the entire PDB file</td>
</tr>
<tr>
<td>JDB Viewer</td>
<td>Link to results of a query of the PDB file from the query tool provided by PDB (includes Medline abstract, crystallization information, secondary structure, protein geometry, options for download and display of the PDB file, molecular visualization and many other relevant links)</td>
</tr>
<tr>
<td>OCA Viewer</td>
<td>Link to results of a query of the PDB file from the OCA Viewer (includes summary of function, residue and secondary structure contacts, structure-derived information such as domain definitions and occurrence of the fold in various genomes, sequence-derived information including links to PDB entry and sequence alignments, molecular visualization and many other relevant links)</td>
</tr>
<tr>
<td>PDBsum</td>
<td>Link to annotation provided by CATH (see below)</td>
</tr>
<tr>
<td>CATH</td>
<td>Link to entry in the CATH database</td>
</tr>
<tr>
<td>Coordinates for external molecular graphics viewer</td>
<td>Molecular viewing using the Rasmol program (running on your local machine)</td>
</tr>
<tr>
<td>WebMol</td>
<td>Simple molecular viewing in your browser using this high-speed Java program</td>
</tr>
<tr>
<td>VRML</td>
<td>Fully featured molecular viewing and analysis using this Java program</td>
</tr>
<tr>
<td>Chimera</td>
<td>Molecular viewing after converting the PDB file to a VRML wireframe model</td>
</tr>
<tr>
<td>3DinSight</td>
<td>Link to entry in this integrated database of biomolecular structure, property and function. Includes protein motifs, graphs of amino acid property, links to PIR and SWISSPROT protein sequence databases and other links.</td>
</tr>
<tr>
<td>Molecules R Us picture</td>
<td>A link to this tool for providing several different static images of the molecule</td>
</tr>
<tr>
<td>PDB</td>
<td>Link to entry in this database of multiple sequence alignments and hidden Markov models</td>
</tr>
<tr>
<td>NCBI Entrez</td>
<td>Link to entry in the molecular modelling database (part of the Entrez sequence retrieval system). Includes link to Medline abstract, sequence and structural neighbours and other links.</td>
</tr>
<tr>
<td>PDB sequence</td>
<td>Displays the protein sequence extracted from the PDB file</td>
</tr>
<tr>
<td>KEGG</td>
<td>Link to entry (if available) in the Kegg encyclopedia of genes and genomes</td>
</tr>
<tr>
<td><strong>CATH (PDBsum)</strong></td>
<td></td>
</tr>
<tr>
<td>Rasmol (icon)</td>
<td>Molecular viewing using the Rasmol program (running on your local machine)</td>
</tr>
<tr>
<td>VRML (icon)</td>
<td>See VRML above</td>
</tr>
<tr>
<td>PDB header</td>
<td>Displays header information from PDB file</td>
</tr>
<tr>
<td>PDB entry (new)</td>
<td>See ‘JDB Viewer’ above</td>
</tr>
<tr>
<td>PDB entry (old)</td>
<td>Link to results of a query of the PDB file from the JDB Browser tool (a precursor to the OCA Viewer, see above)</td>
</tr>
<tr>
<td>PDBsum entry at the NCBI</td>
<td>See ‘NCBI Entrez’ above</td>
</tr>
<tr>
<td>The complete macromolecule</td>
<td>Link to the macromolecular structure database, MSD</td>
</tr>
<tr>
<td>CATH structural classification</td>
<td>Link to entry in the CATH database</td>
</tr>
<tr>
<td>SCOP structural classification</td>
<td>Link to entry in the SCOP database</td>
</tr>
<tr>
<td>FSSP structural alignments</td>
<td>Link to entry in the FSSP database of protein classification</td>
</tr>
<tr>
<td>PROCHECK summary</td>
<td>Link to a summary of the stereochemical quality of the protein structure</td>
</tr>
<tr>
<td>WHMIF check report</td>
<td>Link to the results of this protein verification tool from the WHAT IF program</td>
</tr>
<tr>
<td>PROMOTIF analyses</td>
<td>A link to a summary of the secondary structure and other structural motifs (PDBsum also provides a picture of the secondary structure annotated with residue interactions with the ligand, if appropriate)</td>
</tr>
<tr>
<td>TIPS</td>
<td>Link to a protein topology cartoon</td>
</tr>
<tr>
<td>SAS</td>
<td>Link to an annotated FASTA alignment of sequences in PDB that are related to the protein</td>
</tr>
<tr>
<td>PROSITE</td>
<td>Link to the entry in the PROSITE database of protein patterns (PDBsum also provides a coloured view of the pattern)</td>
</tr>
<tr>
<td>Legend</td>
<td>A summary of the ligand including links to other PDB files containing the same ligand, schematic diagrams, Raster3D picture of the ligand, and a LIGPLOT schematic diagram of protein-ligand interactions.</td>
</tr>
<tr>
<td>MoScript picture</td>
<td>Generate and view a postscript file containing an image of the molecule</td>
</tr>
</tbody>
</table>

© HENRY STEWART PUBLICATIONS (1847-5463). BRIEFINGS IN BIOINFORMATICS. VOL 1, NO 3. 305–315, SEPTEMBER 2000
will associate sequences of unknown structures with CATH superfamilies.

OTHER TOOLS FOR EXPLORING PROTEIN STRUCTURE RELATIONSHIPS

The high quality of the SCOP and CATH databases arises from the way in which experts in protein structure compare proteins manually. However, databases of protein structure classification have been generated by fully automatic means and should not be overlooked. For example, the FSSP database\(^1\) uses the Dali program\(^2\) to perform an all-on-all structural alignment for a set of proteins that is representative of PDB. The resultant similarity values are clustered to generate a ‘fold tree’, a hierarchical classification of the representative set. The classification and alignments are available at the FSSP homepage (see Table 2). It is important to note that, in contrast to SCOP and CATH, FSSP considers whole protein chains rather than structural domains. However, domains have been automatically delineated and classified in the Dali domain dictionary\(^3\), which is derived using Dali and complements FSSP. The dictionary includes a navigable fold tree, a graphical representation of protein fold space, and provides a means for generating multiple alignments and 3D superimpositions for groups of structural neighbours. 3Dee is an alternative database of protein domain definitions in which proteins are clustered on the basis of structural similarity as determined by the STAMP structural alignment program\(^4\). The classification is available on the WWW (see Table 2) as a SCOP-like tree. Tools such as 3dSearch and CE\(^5\) take a PDB file and generate a list of structural neighbours that possess a similar structure to the query protein. The list of neighbours includes appropriate links into the SCOP database allowing easy interpretation of the results of a search. Structural neighbours are also included in the molecular modelling database (MMDDB),\(^6\) which is an integrated part of the Entrez sequence retrieval system.

CONCLUSION

SCOP and CATH are convenient, comprehensive and complementary resources for studying protein structural, functional and possible evolutionary relationships. Protein structure comparison can be used to reveal homologous relationships that are not detectable at the level of sequence. The identification of a common structure can be used to infer functional properties, such as the possible nature of the ligand, mechanisms of ligand binding and so on. However, not all protein relationships are documented in SCOP and CATH so these databases should be augmented by the use of the automatically generated databases of protein structure classification and the various tools for exploring protein structure space. Although a common fold usually points to a common function, there are many cases, particularly the ‘superfolds’, that can be adapted to different protein functions. Therefore, one should always exercise caution when attempting to assign function from a comparison of protein structures.

Acknowledgement

The author gratefully acknowledges the support of the MRC.

References


© ISTE/JOHN WILEY & SONS 2000


