Chapter 2

Protein Structure Databases

Roman A. Laskowski

Abstract

Web-based protein structure databases come in a wide variety of types and levels of information content. Those having the most general interest are the various atlases that describe each experimentally determined protein structure and provide useful links, analyses, and schematic diagrams relating to its 3D structure and biological function. Also of great interest are the databases that classify 3D structures by their folds as these can reveal evolutionary relationships which may be hard to detect from sequence comparison alone. Related to these are the numerous servers that compare folds—particularly useful for newly solved structures, and especially those of unknown function. Beyond these are a vast number of databases for the more specialized user, dealing with specific families, diseases, structural features, and so on.

Key words Protein structure, Protein Data Bank, PDB, wwPDB, JenaLib, OCA, PDBe, PDBsum, ESD, Pfam, CATH, SCOP, Secondary structure, Fold classification, Protein–ligand interactions

1 Introduction

Looking back to 1971, when the Protein Data Bank was founded [1], one cannot help feeling that the study of protein structure must have been a lot simpler then. There were only seven experimentally determined protein structures at the time, and the data for each, including the proteins' atomic coordinates, were stored in simple, fixed-format text files. Admittedly, accessing and displaying this information was trickier, and computers with graphics capabilities tended to be bulky and expensive things. These days, access and display of the data over the Web are vastly easier, but with this comes the problem, not only in the huge increase in the amount of information, but in the multiplicity of sources from which it can be obtained. New servers and services continually appear, while existing ones are modified and improved. Conversely, other servers are abandoned, switched off or neglected, becoming more and more out of date with time. Thus it has become really difficult to know where to go to get relevant answers most easily. Various lists are available on the Web—for example the Nucleic Acids Research

Oliviero Carugo and Frank Eisenhaber (eds.), *Data Mining Techniques for the Life Sciences*, Methods in Molecular Biology, vol. 1415, DOI 10.1007/978-1-4939-3572-7_2, © Springer Science+Business Media New York 2016

(NAR) list at http://www.oxfordjournals.org/our_journals/nar/database/a. This chapter aims to highlight some of the more useful, and up-to-date (at time of writing), sources of information on protein structure that are currently available.

2 Structures and Structural Data

2.1 Terminology

Firstly, it is important to define what is meant by the term "protein structure." It is a term that tends to be somewhat loosely used. A preferable term is "model," as the 3D structures of large molecules such as proteins are models of the atom types, atomic *x*-, *y*-, *z*-coordinates and other parameters that best fit the experimental data. The reason the term "structure" is so commonly used for these models is to distinguish them from "theoretical," or "homology-built," models. Nevertheless, it is important to remember that all are models of reality and that only the former type is based on experimental evidence.

Another loosely used term is "database." Technically, the databases mentioned here are not databases at all, but rather "data resources"—many of which rely on a database for storing and serving up the data. However, the term "database" is becoming common usage for the types of resources described here (e.g., the NAR Database issues), so it is the meaning we adopt here.

2.2 The Protein Data Bank (PDB) and the wwPDB

The primary repository of 3D structural data on proteins (and other biological macromolecules, including RNA, fragments of DNA, carbohydrates, and different complexes of these molecules) is the Protein Data Bank. As mentioned above, this was founded in 1971 and was located at Brookhaven National Laboratories. In October 1998, the management of the archive was taken over by the Research Collaboratory for Structural Bioinformatics (RCSB), a consortium consisting of Rutgers University, the National Institute of Standards and Technology (NIST) and the San Diego Supercomputer Center [2]. Since 2003 the archive has been managed by an international consortium called the world-wide Protein Data Bank (wwPDB) whose partners comprise: the RSCB, the Protein Data Bank Europe (PDBe) at the European Bioinformatics Institute (EBI), the Protein Data Bank Japan (PDBj) at Osaka University, and, more recently, the BioMagResBank (BMRB) at the University of Wisconsin-Madison [3, 4]. Access to the primary data is via the wwPDB's website: http://www.wwpdb.org. The data come in three different formats: old-style PDB-format files, macromolecular Crystallographic Information File (mmCIF) format [5], and a XML-style format called PDBML/XML [6]. Due to format limitations, the old-style PDB-format files are no longer available for extremely large structural models (i.e., those having too many atoms, residues or chains than the fixed-format fields allow for). For many of the structures, the wwPDB also make the original experimental data available. Thus, for structural models solved by X-ray crystallography, one can often download the structure factors from which the model was derived, while for structures solved by nuclear magnetic resonance (NMR) spectroscopy, the original distance and angle restraints can be obtained. As of July 2015, the wwPDB contained over 110,000 structural models, each identified by a unique 4-character reference code, or PDB identifier.

A key task the wwPDB have performed is the remediation of the legacy PDB archive to fix and make consistent the entire PDB data, in particular relating to ligands and literature references [7]. The PDBe and UniProt groups at the EBI have mapped the sequences in the PDB entries onto the appropriate sequences in UniProt [8]. More recently, the focus has been on validation of the structural data, with the establishment of several Validation Task Forces [9–11], and the reporting of quality indices or validation information for each structure.

2.3 Structural Data and Analyses

Rather than download the raw data from the wwPDB for each protein of interest, it is usually more convenient to obtain the required information directly from one of the myriad protein structure databases on the Web. These come in many shapes and sizes, catering for a variety of needs and interests.

At the simplest level are the sites that provide “atlas” pages—one for every PDB entry—each containing general information obtained from the relevant PDB file. There are usually graphical representations of the structural model together with links that provide interactive 3D visualizations using Java-based, or other, viewers. Each of the founding members of the wwPDB have their own atlas pages: the RCSB, the PDBe, and PDBj. In addition, there are several other sites that have much to commend them, and some of these are mentioned below.

Beyond the atlases, there are a host of other types of sites and servers. These include those providing information on specific structural motifs, focus on selected protein families, classify protein folds, compare protein structures, provide homology-built models for proteins for which no structure has been determined, and so on. This chapter cherry-picks a few of the more interesting and useful sites to visit.

3 Atlases

Table 1 lists the seven best-known and useful of the atlas sites. All have been developed independently and, not unexpectedly, all have much in common as the information comes from the same source: the PDB entry. The protein name, authors, key reference,

Table 1
Protein structure atlases

Server	Location	URL	References
JenaLib	Fritz Lipmann Institute, Jena, Germany	jenalib.fli-leibniz.de/	[30]
MMDB	NCBI, USA	www.ncbi.nlm.nih.gov/Structure/MMDB/mmdb.shtml	[55]
OCA	Weizmann Institute, Israel	oca.weizmann.ac.il	[56]
PDBe	EBI, Cambridge, UK	www.ebi.ac.uk/pdbe	[26]
PDBj	Osaka University, Japan	www.pdbj.org	[57]
PDBsum	EBI, Cambridge, UK	www.ebi.ac.uk/pdbsum	[32, 33]
RCSB	Rutgers and San Diego, USA	www.rcsb.org/pdb	[18]

experimental methods, atomic coordinates, and so on are obviously identical on all sites. Also common are certain derived data, including quality assessment of each structural model, and information about the protein's likely "biological unit."

Quality assessment is a crucial issue as not all models are equally reliable, and much has been written on this topic over the years [9, 12–16]. The main problem is that the results of any experiment contain errors, but for structural models it is difficult to estimate the extent of those errors. For X-ray models, a rough guide of quality is provided by the resolution at which the structure was solved and its *R*-factor, but for NMR models there is no such ready measure. Some atlases do provide indications of which models are more reliable, as described shortly.

The second important issue is knowing what a given protein's biological unit is. This is not always obvious from the PDB entry. The problem is that the deposited coordinates from an X-ray crystal structure determination correspond to the molecule(s) in the asymmetric unit. This may give a false impression of how the protein operates in vivo. For example, what may look like a monomer from the PDB entry, is, in real life, a dimer, or a trimer, etc. Conversely, the PDB entry might give the coordinates of a dimer, yet the biological unit happens to be a monomer—the dimeric structure being the result of packing in the crystal. For any structural analysis it is crucial to know what the true biological unit is. For some proteins the biological unit has been determined experimentally, and so is known with great confidence. In others it has to be deduced computationally by analysis of the packing of the individual chains in the crystal. Some interfaces are more substantial than others and hence likely to represent genuine biological interactions rather than happenstance crystal contacts. Most of the atlases provide information on the known, or predicted, biological unit. The most commonly used prediction method is Protein Interfaces, Surfaces and Assemblies (PISA) [17].

Beyond these general similarities, the atlases differ in sufficient respects to make them complement one another; they differ in what additional information they pull in, the links they make to external resources, and the analyses of the 3D structures they provide. Consequently, the atlas of choice can be either a matter of personal preference or depend on the type of information required at the time.

Here we focus only those aspects that make each one unique, useful or interesting. We start with the atlases provided by the founding members of the wwPDB, and then discuss some of the others.

3.1 The RCSB PDB

The RCSB's website [18] has been revamped several times and is an extremely rich source of information about each PDB entry. It used to be a little overwhelming for novices, but recently a great deal of effort has gone into simplifying the design as well as adding new information—such as the relationship of structures to their corresponding genes and to associated diseases and therapeutic drugs. A specific aim of the website has been to "bring a structural view of biology and medicine to a general audience."

3.1.1 Summary Page

Figure 1 shows the summary page for PDB entry 1ayy, a glycosylasparaginase. The top box shows the primary citation for this entry, being the published description of the experiment that resulted in the structural model and any analysis the authors might have performed on it, including relating the structure to the protein's biological function. To the right is a thumbnail image of the protein and links for viewing it in one of three molecular graphics viewers. The "More Images" link shows the asymmetric unit and the biological unit, as described above (although in many cases they are identical). The latter is either as defined by the depositors or as predicted by the PISA algorithm.

The Molecular Description box provides a schematic diagram of the protein's sequence and structural domains, together with its secondary structure, and which parts of the protein the structure corresponds to. An expanded view can be obtained by clicking on "Protein Feature View," as shown in Fig. 2. Often structural models are not of the whole protein but merely cover one or two domains or, in some cases, are mere fragments of the protein. The diagram makes it clear what the coverage is. The little plus symbol at the bottom opens up a window showing other known structures of the same protein—which is particularly useful in identifying structures that may be more complete, or solved at a higher resolution. The sequence domains are as defined by Pfam [19], while the structural domain definitions come from SCOP [20].

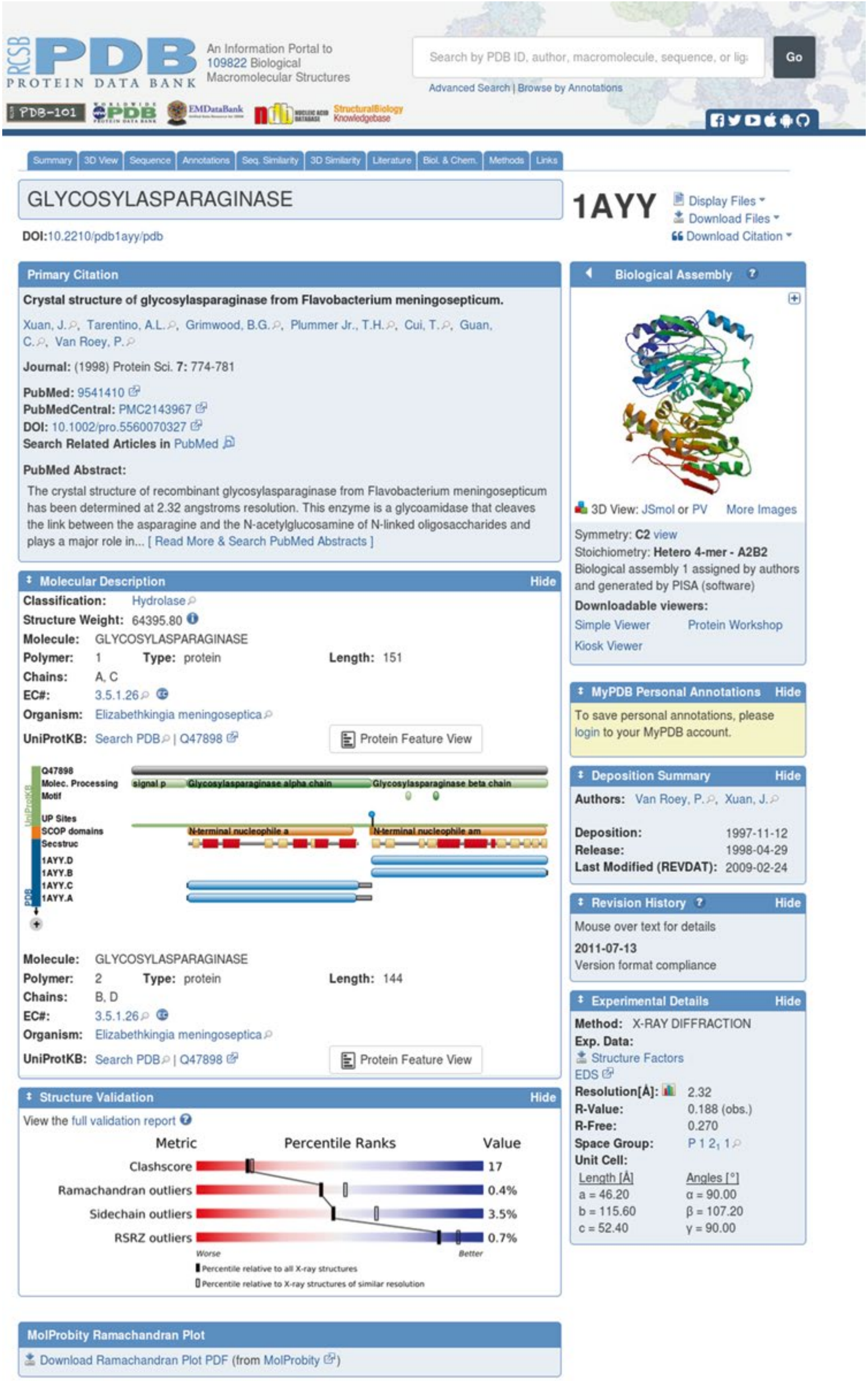

Fig. 1 Part of the RCSB atlas page for PDB entry 1ayy, a glycosylasparaginase determined by X-ray crystallography at 2.32 Å

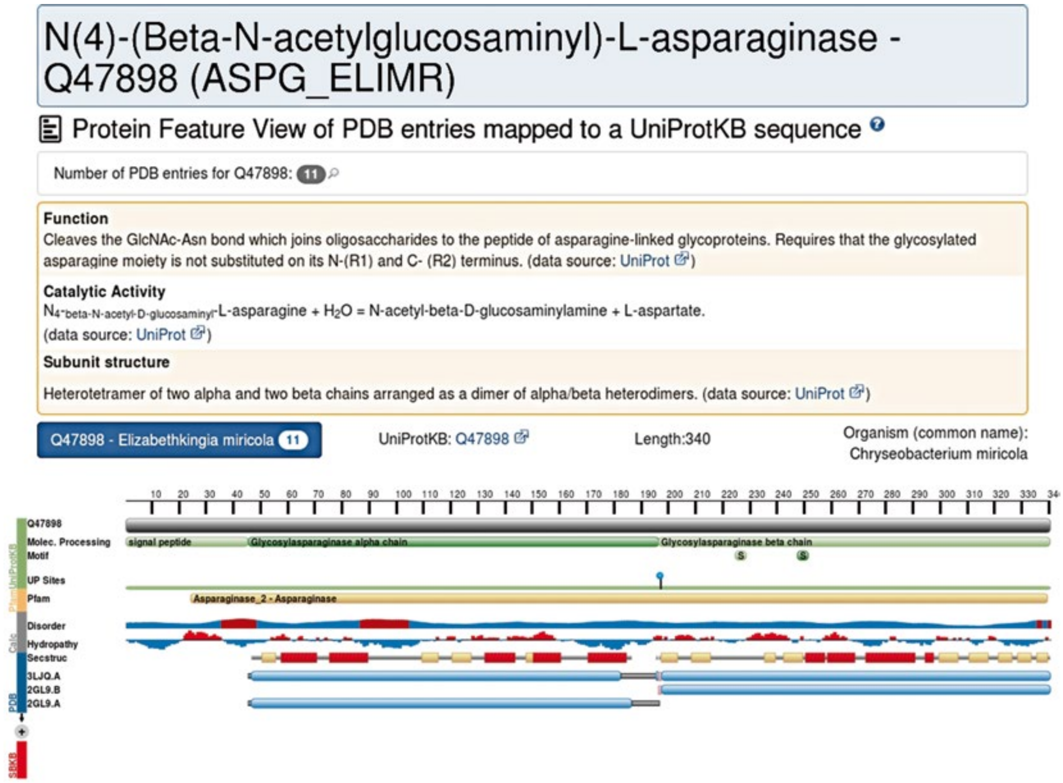

Fig. 2 The Protein Feature View of PDB entry 1ayy on the RCSB PDB server. The diagram shows the protein's sequence (Pfam) and structural (SCOP) domains, its hydrophathy, secondary structure, and structural coverage

The large box at the bottom of Fig. 1 is a "validation report slider" providing an at-a-glance assessment of the structure's likely quality (only available for X-ray models). The graphic indicates how the structure compares on a number of quality-related parameters against all other structures in the database as well as structures solved at the same resolution. The parameters include the R_{free}, an atom-atom "clash score," number of Ramachandran plot outliers as computed by the MolProbity structure validation program [21], and the real-space R-value Z-score as computed by the Uppsala Electron-Density Server [22]. An almost identical schematic is provided by the PDBe website (*see* Fig. 3). A link above the schematic provides the full validation report for the structure in question.

3.1.2 Other Information

Besides the summary information, further structural details are presented on additional pages titled: 3D View, Sequence, Annotations, Seq. Similarity, 3D Similarity, Literature, Biology & Chemistry, Methods, and Links.

For ligands there is the 3D Java-based Ligand Explorer [23] which allows you to select and view different types of protein–ligand interactions. There is also a schematic 2D PoseView [24] diagram of the protein–ligand interactions.

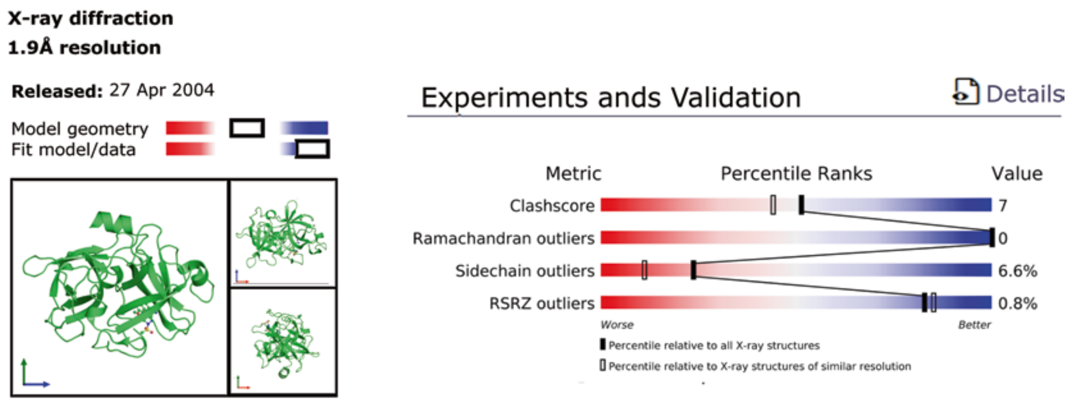

Fig. 3 Validation schematics for PDB entry 1sqt, as shown on the PDBe website. Above the thumbnail images of the protein on the left are two "quality sliders." The *top one* shows how well the overall model quality compares against all other structures in the PDB, and the second how well the model fits the experimental data from which it was derived. The *red end* of the slider indicates a poor model/fit, while the *blue* indicates the model is a good one. The *right-hand set* of sliders show the quality of the model as judged by four different global quality criteria: the R_{free}, an atom-atom clash score computed by MolProbity, number of Ramachandran plot outliers, and the real-space *R*-value *Z*-score as computed by the Uppsala Electron-Density Server. The *black vertical box* on each slider corresponds to the percentile rank of the given score with respect to the scores of previously deposited PDB entries, while the *white vertical box* shows the rank with respect to entries solved at a similar resolution

The advanced search option allows for quite complex queries and subqueries on the data, telling you how many hits each set of conditions returns as you refine your search.

3.1.3 Molecule of the Month

One particularly eye-catching feature of the RCSB site is the "Molecule of the Month" pages written by David S. Goodsell of The Scripps Research Institute and illustrated with his beautiful plots [25]. Each month the structure and function of a different protein or protein family is described, with specific references to the PDB entries that have contributed to the understanding of how the proteins achieve their biological roles. The collection of short articles, which are suitable for specialists and non-specialists alike, dates back to the year 2000 and now numbers over 180 entries, providing a nice reference and educational resource. Additionally, and particularly useful as teaching materials, are the accompanying videos, posters, lesson plans and curricula provided by the PDB-101 educational portal.

3.2 The PDBe

The website of the Protein Data Bank Europe (PDBe) [26] has many similarities to the RCSB's. The atlas pages for each entry show the usual summary information describing the structure and the experimental details used to obtain it. Additional pages relate to Structure analysis, Function and Biology, Ligands and Environments, and Experiments and Validation. The Molecular

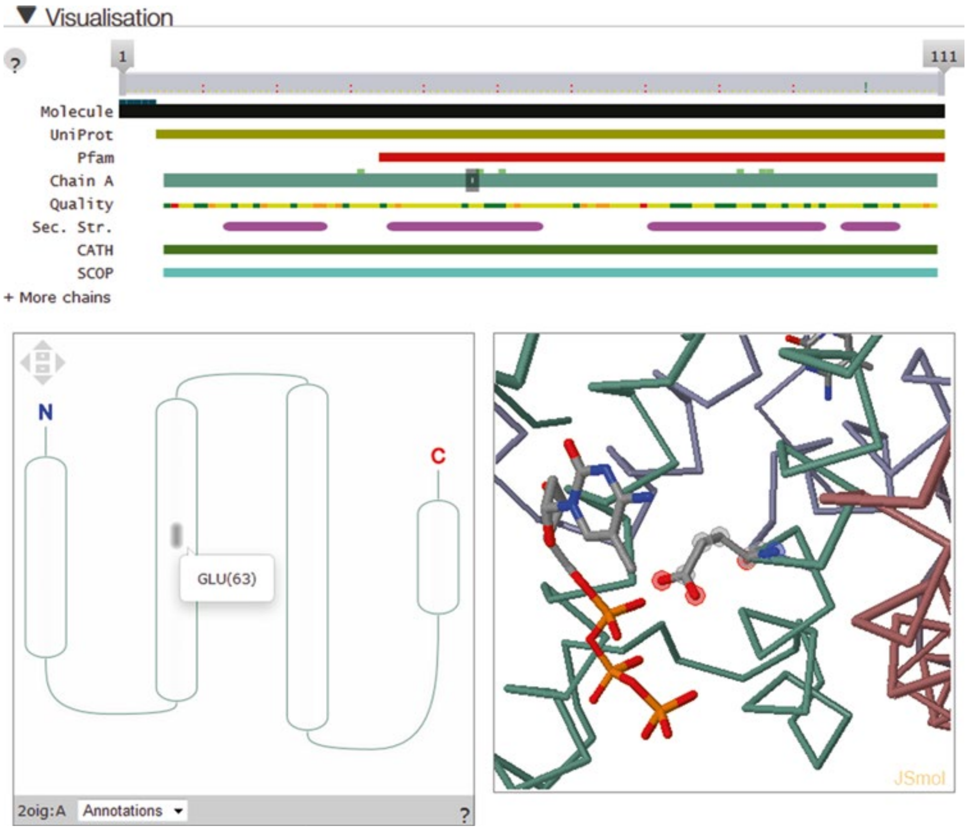

Fig. 4 The "Molecule details" of PDB entry 2oig, a mouse dCTP pyrophosphatase 1, from the PDBe website. The tracks at the top represent the protein's sequence and structure domains, its secondary structure and residue-by-residue quality indicators. It is similar to the RCSB's Protein Feature View in Fig. 2. At *bottom left* is a topology diagram of the secondary structure elements—here four helices. Clicking on the diagram identifies the residues, and the corresponding residues are highlighted in the diagram above (by a *shaded grey box*) and in the JSmol 3D image on the right

Details link shows the protein's sequence features, a diagram of its secondary structure topology and a 3D JSmol view (Fig. 4). These are connected such that clicking on one diagram highlights the corresponding residues in the others.

3.2.1 PDBeFold

In addition to the atlas pages, the PDBe website has a number of useful applications. These include PDBeFold which performs fold matching of any one or more protein structures against one or more others. The server makes use of the secondary structure similarity matching program SSM [27]. You can match a single PDB entry against another, or against all structures in the PDB. You can upload your own PDB-format file, or a list of PDB pairs to compare. The outputs include structure-based alignments with computed rmds values and various scores of significance. The superposed structures can be viewed or their coordinates downloaded.

3.2.2 PDBeMotif

PDBeMotif [28, 29] allows searches for sequence and structural motifs as well as for ligands and specific protein–ligand interactions. Structural motifs can be defined in terms of patterns of secondary structure, φ/ψ and χ angles, and C^{α} and side-chain positions. Searches are entered either via a simple Web form or using a graphical query generator. The hits from a search can be viewed in three dimensions, aligned by ligand, PROSITE pattern, active site residues or by environment. One can generate various statistics on protein–ligand interactions (e.g., to compare the different distributions of residues binding to ATP and GTP). Of particular use is an option to upload a PDB file and scan its ligands and active sites against the PDBe data.

3.2.3 PDBePISA

PDBePISA is a service for computing the stability of protein–protein or other macromolecular complexes (protein, ligands, and DNA/RNA). It uses the PISA [17] program and provides an analysis of the surfaces, interfaces, and assemblies to suggest which groupings are likely to be biological assemblies rather than crystal packing ones. The assessment is based on the number, type, and strength of interactions across each interface. The service is especially useful for obtaining the full biological units for large multimeric complexes where the PDB entry consists only of a single protein chain.

3.3 JenaLib

The Jena Library of Biological Macromolecules, JenaLib [30], was one of the earliest sites offering atlas pages for each PDB entry, specializing in hand-curated images of the structures showing functionally informative views. Rather than split information across several pages, JenaLib shows all the information on a single page but has a collapse/expand mechanism for controlling what is shown and what is hidden. In addition to several of the standard 3D viewers the site features its own: the JenLib Jmol viewer. This viewer is an extension of Jmol which has a number of options not found in other viewers, such as highlighting of PROSITE motifs, single amino acid polymorphisms and CATH [31] or SCOP domain structures.

JenaLib has more links to external databases than the other atlas sites and is particularly strong on its many visualizations of each entry—both in terms of its interactive viewing options and its preprepared still images.

A particularly useful feature is a form for generating lists of PDB entries according to a number of criteria. Additionally, there are a number of precomputed lists of structures; for example, all nucleic acid structures without protein, all carbohydrate structures, and so on.

3.4 OCA

OCA's main difference from the other atlases is its linkage between proteins and the diseases associated with them. It differs also in that its home page is a search form, with searches possible on gene name, function, disease and membrane orientation (for membrane-spanning proteins).

3.5 PDBsum

The last of the atlases described here is PDBsum [32, 33]. Its original aim was to provide pictorial structural analyses where other sites were presenting tables of numbers, but the other atlases have come to include more schematic diagrams over the years. It still provides some unique features, including an option that allows users to upload their own PDB files and get a set of password-protected PDBsum pages generated for them.

3.5.1 Summary Page

Each entry's summary page has a thumbnail image of the structure, the usual header information and a clickable schematic diagram showing how much of the full-length protein sequence is actually represented by the 3D structural model. The diagram shows the protein's secondary structure and annotates it with any Pfam sequence domains and CATH structural domains. Also included is a thumbnail Ramachandran plot of the protein and the primary citation.

3.5.2 Quality Assessment

Hovering the mouse over the thumbnail Ramachandran pops up a full-size version. A reliable model will have more points in the core regions (colored red) and, ideally, none in the cream-colored, disallowed regions. Residues in the latter are labeled, so if a model has many labeled residues, it might be an idea to look for an alternative. Clicking on the plot goes to a page showing the summary results from the PROCHECK quality assessment program [34] and from this page you can generate a full PROCHECK report.

3.5.3 Enzyme Reactions

For enzymes, the relevant reaction catalyzed by the enzyme is shown by a reaction diagram where possible. If any of the ligands bound to the protein correspond to any of the reactants, cofactors or products, the corresponding molecule in the diagram is boxed in red. If a ligand is merely similar to one of these, a blue box surrounds the molecule instead and a percentage similarity is quoted.

3.5.4 Figures from Key references

The majority of experimentally determined protein structures are reported in the scientific literature, often in high profile journals, and each PDB file cites the "key" reference—i.e., the one describing the structure determination, analysis and biological significance of the protein. Like the other atlas sites, PDBsum cites this reference, shows its abstract and provides links to both the PubMed entry and to the online version of the article. Where PDBsum differs is that for many of these references it also gives one or two figures (plus figure legends) taken directly from the key reference itself [35]. This is done with permission from the relevant publishers and is useful for two reasons. Firstly, a carefully selected figure can speak volumes about an important aspect of the protein's structure or function. And secondly, each paper's lead author is requested to review which figures have been selected by the automated process and, if need be, suggest better choices. About one in six authors take the trouble to do this. And some even add an additional comment to appear on the entry's summary page (e.g., PDB entry 1hz0).

3.5.5 Secondary Structure and Topology Diagrams

From the summary page are various additional pages giving schematic diagrams of different aspects of the 3D structure. The "Protein" page shows a diagram of the chain's secondary structure elements, much like the RCSB's diagram shown in Fig. 2. Additional features include the annotation of residues that are catalytic—as defined in the Catalytic Site Atlas (CSA) [36]—or are included in the SITE records of the PDB file, or interact with a ligand, DNA/RNA or metal, or belong to a PROSITE pattern [37]. CATH structural domains are marked on the sequence, in contrast to the RCSB's diagram which uses SCOP. Where there is information on the conservation of each residue in the sequence—obtained from the ConSurf-HSSP site [38]—the secondary structure plot can be redisplayed with the residues colored by their conservation.

Next to the secondary structure plot is a topology diagram either of the whole chain or, where it has been divided into its constituent CATH domains, of each domain (Fig. 5). The diagram shows the connectivity of the secondary structure elements, with the constituent β-strands of each β-sheet laid side-by-side, parallel or antiparallel, to show how each sheet in the chain/domain is formed, and where any helices are found relative to the sheets.

3.5.6 Intermolecular Interactions

Some of the other pages are devoted to schematic representations of intermolecular interactions. Thus for each ligand molecule or metal ion in the structure there is a schematic LIGPLOT diagram [39] of the hydrogen bonds and non-bonded interactions between it and the residues of the protein to which it is bound (*see* Fig. 6). Similarly, any DNA–protein interactions are schematically depicted by a NUCPLOT diagram [40]. Protein–protein interactions at the interface between two or more chains are shown by two plots: the first shows an overview of which chains interact with which (Fig. 7b), while the second shows which residues actually interact across the interface (Fig. 7c).

4 Homology Models and Obsolete Entries

4.1 Homology Modeling Servers

As mentioned above, there were over 110,000 structural models in the wwPDB as of July 2015. However, some were not of proteins and many were duplicates: that is the same protein solved under different conditions, or with different ligands bound, or with one

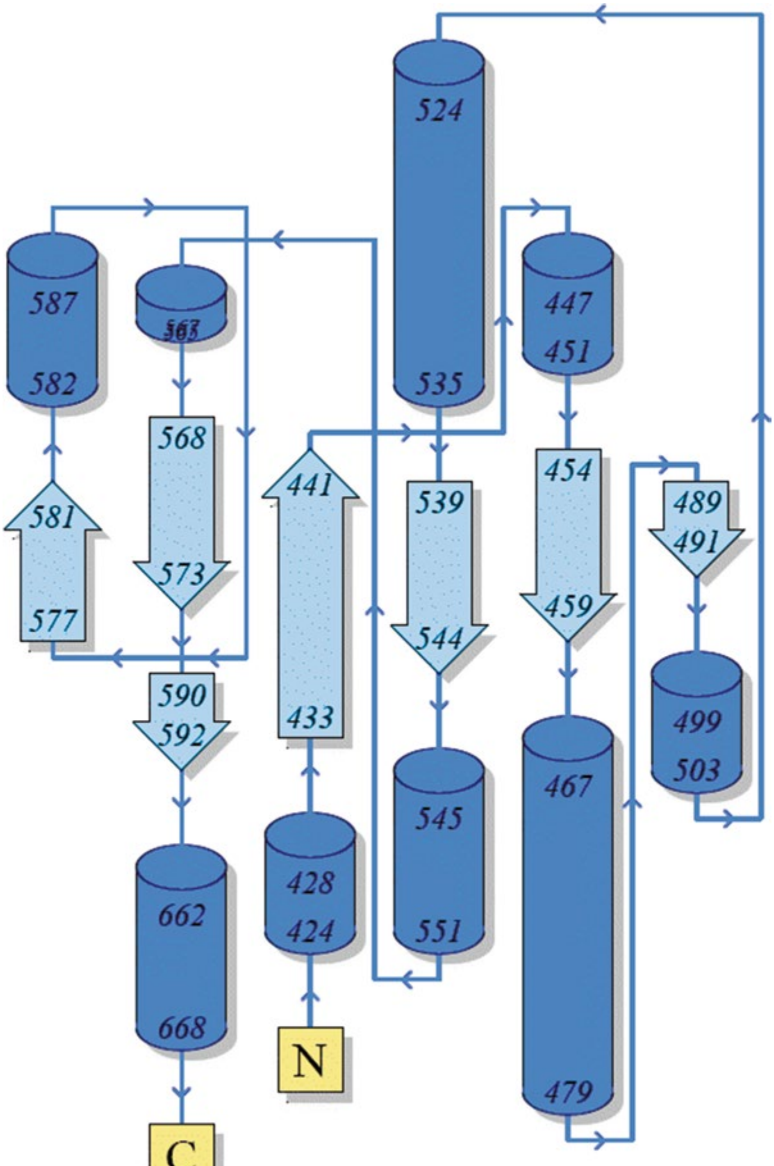

Fig. 5 A topology diagram taken from PDBsum for the second domain of chain A in PDB entry 2b6d: a bovine lactoferrin. The diagram illustrates how the β-strands, represented by the *block arrows*, join up, side-by-side, to form the domain's central β-sheet. The diagram also shows the relative locations of the α-helices, here represented by cylinders. The small *arrows* indicate the directionality of the protein chain, from the N- to the C-terminus. The *numbers* within the secondary structural elements correspond to the residue numbering given in the PDB file

or more point mutations. In terms of unique protein sequences, as defined by the UniProt identifier, this 110,000 corresponded to only about 33,000 unique proteins. (Compare this number with the 620 million protein sequences in the European Nucleotide

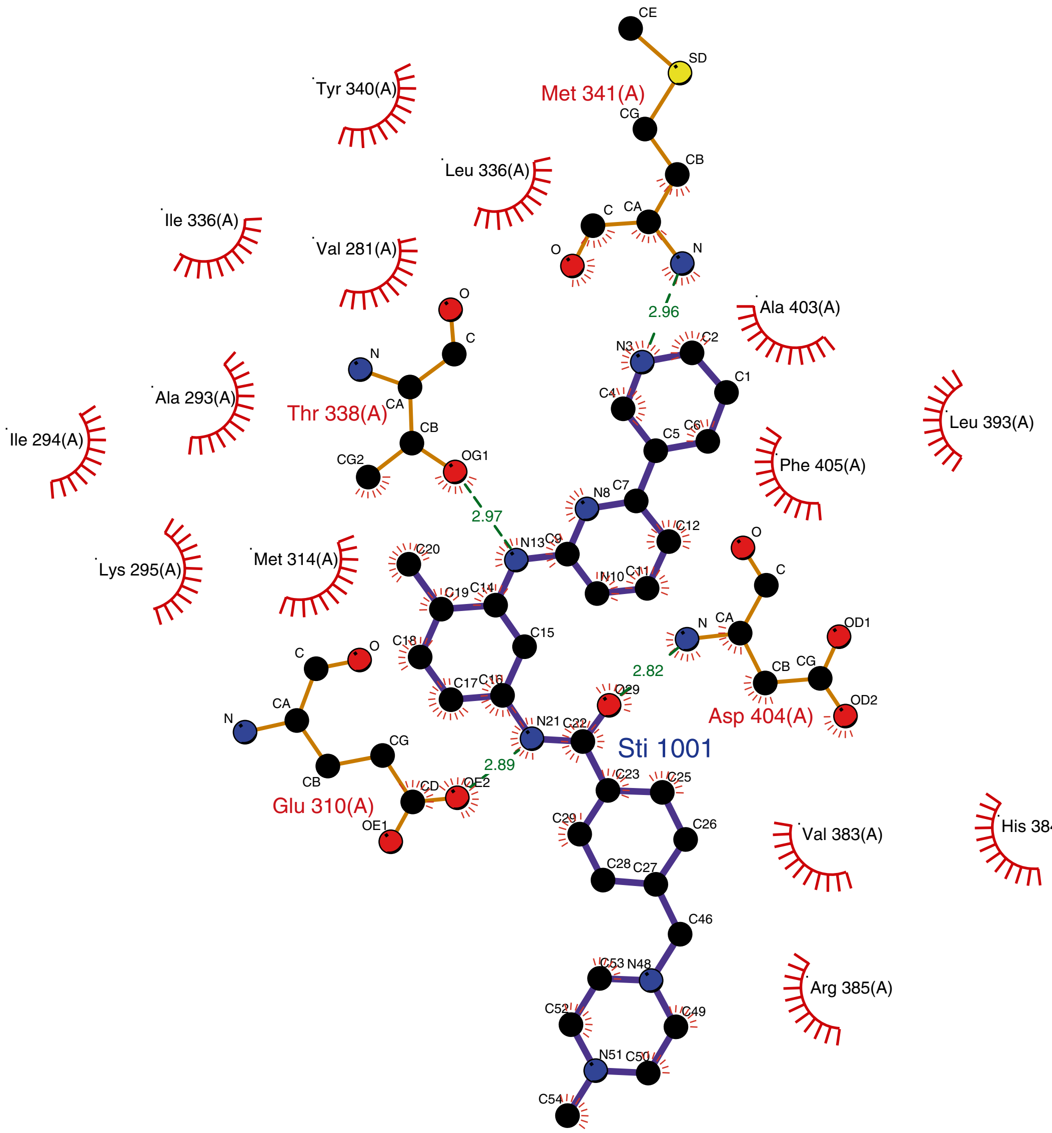

Fig. 6 LIGPLOT for PDB entry 2oig, tyrosine kinase c-Src, as given in PDBsum showing the interactions between the bound molecule imatinib (a drug, brand name Gleevec) with the residues of the protein. Hydrogen bonds are represented by *dashed lines*. Residues that interact with the ligand via non-bonded contacts only are represented by the eyelashes

Archive (ENA) [41]). Moreover, for many of these, the 3D structure represents only a part of the full sequence—a single domain or just a fragment.

So for many proteins there is no corresponding structural model in the PDB. In these cases it is common to build a homology model based on the 3D structural model of a closely related protein (if there is one). The PDB used to accept homology-built models together with the experimentally determined ones but, as

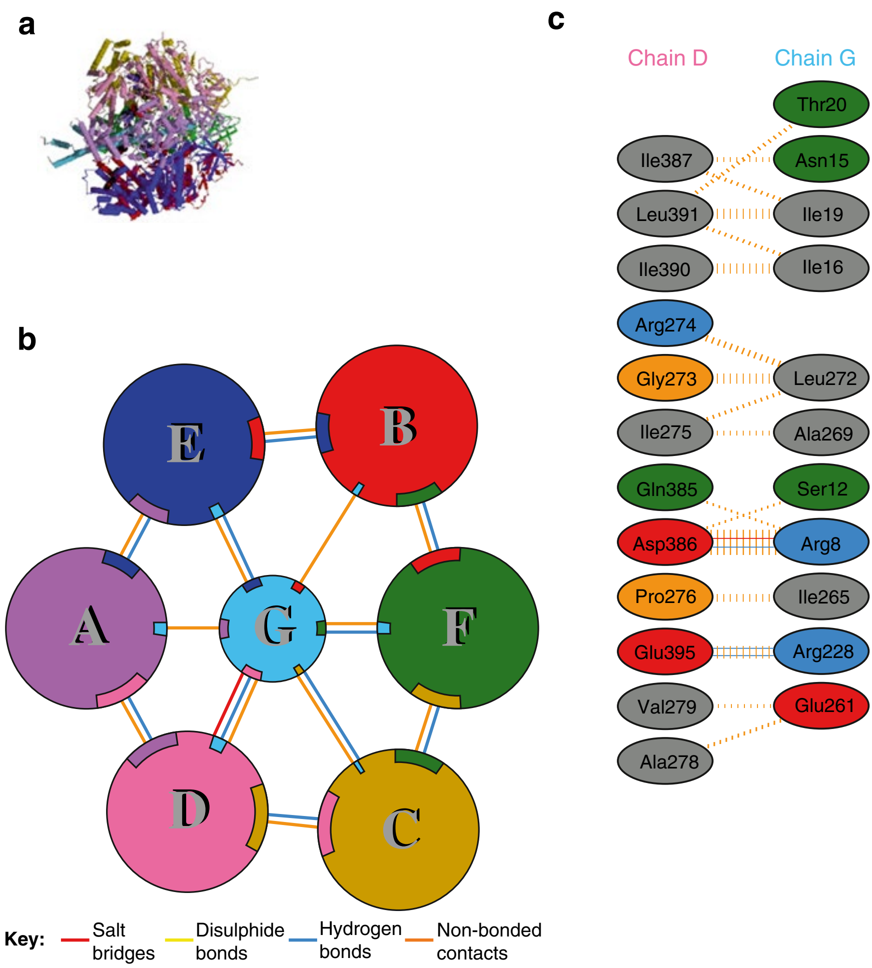

Fig. 7 Extracts from the protein–protein interaction diagrams in PDBsum for PDB entry 1cow, bovine mitochondrial F1-ATPase. (**a**) Thumbnail image of the 3D structural model which contains seven protein chains: three of ATPA1_BOVIN (chains A, B, and C), three of ATPB_BOVIN (chains D, E, and F), and a fragment of ATPG_BOVIN (chain G). (**b**) Schematic diagram showing the interactions between the chains. The area of each circle is proportional to the surface area of the corresponding protein chain. The extent of the interface region on each chain is represented by a colored wedge whose color corresponds to the color of the other chain and whose size signifies the interface surface area. (**c**) A schematic diagram showing the residue–residue interactions across one of the interfaces, namely that between chains D and G. Hydrogen bonds and salt bridges are shown as *solid lines* while non-bonded contacts are represented by *dashed lines*

of 1 July 2002, moved its holding of theoretical models out of the standard PDB archive to a separate ftp site and then, as of October 15, 2006, stopped accepting any new ones. As of July 2015 there were only 1358 models on the ftp site so, with such a small number, it is unlikely that one's protein of interest will be among them.

The alternative is to build a homology model oneself, and there are various servers that will perform the process largely, or completely, automatically. The best-known is SWISS-MODEL [42]. This accepts a protein sequence and will return a 3D model if it is able to build one. More advanced users can submit multiple sequence alignments and manually refine the final model. It is important to remember that any homology-built model will, at best, be imperfect and at worst totally misleading—particularly if one or more of the structural models that act as a template for the model contain errors. So a key part of SWISS-MODEL are the various validation checks applied to each model to provide the user with an idea of its likely quality.

Table 2i shows a list of automated homology modeling Web servers.

Aside from building a model yourself, it may be possible to download a ready-built, off-the-shelf one. The SWISS-MODEL Repository [43] contained over three million models in July 2015, each accessible by its UniProt accession number or identifier. Similarly ModBase [44] contains a large number of precomputed models for sequences in the SwissProt and TrEMBL databases—34 million models for 5.7 million proteins in July 2015. Table 2iii gives the URLs and references for these servers.

Table 2
Homology model servers

Server	Location	URL	References
i. Automatic homology modeling			
3D-JIGSAW	Imperial Cancer Research Fund, UK	bmm.cancerresearchuk.org/~3djigsaw	[58]
CPHmodels	Technical University of Denmark	www.cbs.dtu.dk/services/CPHmodels	[59]
ESyPred3D	University of Namur, Belgium	www.fundp.ac.be/urbm/bioinfo/esypred	[60]
SWISS-MODEL	Biozentrum Basel, Switzerland	Swissmodel.expasy.org	[42]
ii. Evaluation of modeling servers			
CAMEO	Swiss Institute of Bioinformatics and Biozentrum Basel, Switzerland	www.cameo3d.org/	[61]
iii. Precomputed homology models			
SWISS-MODEL Repository	Biozentrum Basel, Switzerland	Swissmodel.expasy.org/repository	[43]
ModBase	University of California San Francisco, USA	modbase.compbio.ucsf.edu	[44]
PDB archive	RCSB, USA	ftp://ftp.rcsb.org/pub/pdb/data/structures/models	

4.2 Threading Servers

What if there is no sufficiently similar protein of known structure and thus no possibility of building a homology model? In these cases, it is sometimes necessary to resort to desperate measures such as secondary structure prediction and fold recognition, or "threading." The results from these methods need to be treated with extreme care. Occasionally, these methods approximate the right answer—usually for small, single-domain proteins where they may produce topologically near correct models [45]—and they are improving all the time [45], but perhaps should only be used only as a last resort.

4.3 Obsolete Entries

As experimental methods improve, better data sets are collected or earlier errors are detected, so some structural models in the PDB become obsolete. Many are replaced by improved structural models, whereas others are simply quietly withdrawn. None of these obsolete entries disappear entirely, though. Some of the atlases mentioned above include the obsolete entries together with the current ones. The RCSB website provides a full list at: http://www.rcsb.org/pdb/home/obs.do.

5 Fold Databases

5.1 Classification Schemes

In 2006, it was estimated that there are around 900 known fold groups [46]. Many proteins comprise more than one structural domain, with each domain being described by its own fold and often able to fold independently of the rest of the protein. There have been a number of efforts to classify protein domains in a hierarchical manner. The two current market leaders in this field are the SCOP and CATH hierarchical classification systems (*see* Table 3i). In CATH, protein structures are classified using a combination of automated and manual procedures, with four major levels in the hierarchy: Class, Architecture, Topology (fold family) and Homologous superfamily [31, 47]. In SCOP the classification is more manual, although some automated methods are employed. Comparisons between the two classification schemes have shown there to be much in common, although there are differences, primarily in how the structures are chopped into domains [48].

However, it appears that protein folds are not the discrete units that these classification schemes might imply, but rather that protein structure space is a continuum [49] and folds can lose core element by a process of "domain atrophy" [50]. Nevertheless, the two databases are very valuable resources because they group domains by their evolutionary relationships even where this is not apparent from any similarities in the sequences.

Table 3
Fold classification and comparison servers

Server	Location	URL	References
i. Automatic homology modeling			
CATH	University College London, UK	www.cathdb.info	[62]
SCOP2	University of Cambridge, UK	scop2.mrc-lmb.cam.ac.uk/	[63]
ii. Fold comparison			
RCSB PDB Protein Comparison Tool	RCSB, USA	www.rcsb.org/pdb/workbench/workbench.do	[64]
Dali	University of Helsinki, Finland	ekhidna.biocenter.helsinki.fi/dali_server	[65]
DBAli	University of California San Francisco, USA	www.salilab.org/DBAli/	[66]
MATRAS	Nara Institute of Science and Technology, Japan	strcomp.protein.osaka-u.ac.jp/matras	[67]
PDBeFold	European Bioinformatics Institute, UK	www.ebi.ac.uk/msd-srv/ssm	[27]
TOPSCAN	University College London, UK	www.bioinf.org.uk/topscan	[68]
VAST+	NCBI, USA	www.ncbi.nlm.nih.gov/Structure/vastplus/vastplus.cgi	[55]

5.2 Fold Comparison

Often a given structural domain is associated with a specific biological function. However, the so-called superfolds, which are more common than other folds, tend to be responsible for a wide range of functions [51]. There are a large number of Web servers, such as PDBeFold mentioned above, that can identify all proteins sharing a given protein's fold. Each server has a different algorithm or a different way of assessing the significance of a match. Table 3ii lists a selection of the more popular servers. A fuller list, together with brief descriptions of the algorithms and a comparison between them, can be found in various comparisons that have been made between them [52, 53].

6 Miscellaneous Databases

6.1 Selection of Data Sets

For any bioinformatics analysis involving 3D structural models it is important to get a valid and representative data set of models of as high a quality as possible. To help in this process there are various servers that allow you to obtain such lists based on various selection criteria. Table 4 lists several such servers.

Table 4
Selection of data sets

Server	Location	URL	References
ASTRAL	University of Berkeley, USA	scop.berkeley.edu/astral	[69]
JenaLib (Entry Lists)	Fritz Lipmann Institute, Jena, Germany	jenalib.fli-leibniz.de/	
PISCES	Fox Chase Cancer Center, Philadelphia, USA	dunbrack.fccc.edu/PISCES.php	[70]

6.2 Uppsala Electron Density Server (EDS) and PDB_REDO

As has been mentioned a couple of times already, a key aspect of any structural model is how reliably it represents the protein in question. A poor quality model limits what structural or functional conclusions can be drawn from it. For X-ray models, in addition to the geometrical checks mentioned in passing above, the most useful guide to reliability is how well the model agrees with the experimental data on which it was based. The Uppsala Electron Density Server, EDS [22], displays the electron density maps for PDB entries for which the experimental structure factors are available. The server also provides various useful statistics about the models. For example, the plots of the real-space *R*-factor (RSR) indicate how well each residue fits its electron density; any tall red spikes are regions to be wary of. Other useful plots include: the occupancy-weighted average temperature factor and a *Z*-score associated with the residue's RSR for the given resolution. The latter is used in the wwPDB's quality slider (*see* Fig. 3).

The above calculations require the original experimental data. Another use for the data is to rerefine the structural models. As refinement methods and software improve, so it is possible to revisit structural models solved in the past and rerefine them to, possibly, get better models. A server devoted to such improvement is PDB_REDO [54] (http://www.cmbi.ru.nl/pdb_redo). This provides validation measures before and after the new refinement showing the degree of improvement of the model.

6.3 Curiosities

Finally, there are various sites which deal with slightly more offbeat aspects of protein structure. Some are included in Table 5. A couple detect knots in protein folds: Protein Knots and the pKnot Web server. The former lists 44 PDB entries containing knotted proteins, classified according to the type of knot. Another interesting site, which can while away part of an afternoon, is the Database of Macromolecular Movement which holds many movies showing proteins in motion. Also included is a "Morph Server" which will produce 2D and 3D animations by interpolating between two submitted protein conformations—very useful for producing animations for presentations or websites.

Table 5
Miscellaneous servers

Server	Location	URL	References
3D Complex	MRC, Cambridge, UK	www.3dcomplex.org/	[71]
Database of Macromolecular Movements	Yale, USA	molmovdb.org	[72]
Electron Density Server (EDS)	Uppsala, Sweden	eds.bmc.uu.se/eds	[22]
Orientations of Proteins in Membranes (OPM)	University of Michigan, USA	opm.phar.umich.edu	[73]
pKnot server	National Chiao Tung University, Taiwan	pknot.life.nctu.edu.tw	[74]
Protein Knots	Massachusetts Institute of Technology, USA	knots.mit.edu	[75]

7 Summary

This chapter describes some of the more generally useful protein structure databases. There are many, many more that are not mentioned. Some are very small and specialized, such as the so-called "hobby" databases, created by a single researcher and lovingly crafted and conscientiously updated—until, that is, the funding runs out, or the researcher moves on to another post and the database is abandoned and neglected. The larger and more widely used databases have better resources to keep them ticking over, but tend to suffer from a great deal of duplication and overlap. This can be seen in the large numbers of PDB atlases and fold comparison servers. Perhaps one day, a single server of each type will emerge combining the finer aspects of all others to make life a lot easier for the end users of the data.

Acknowledgments

The author would like to thank Tom Oldfield for useful comments on this chapter.

References

1. Bernstein FC, Koetzle TF, Williams GJ et al (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. J Mol Biol 112:535–542
2. Berman HM, Westbrook J, Feng Z et al (2000) The Protein Data Bank. Nucleic Acids Res 28:235–242
3. Berman H, Henrick K, Nakamura H (2003) Announcing the worldwide Protein Data Bank. Nat Struct Biol 10:980
4. Berman HM, Kleywegt GJ, Nakamura H, Markley JL (2012) The future of the protein data bank. Biopolymers 99:218–222

5. Westbrook JD, Fitzgerald PM (2003) The PDB format, mmCIF, and other data formats. Methods Biochem Anal 44:161–179
6. Westbrook J, Ito N, Nakamura H, Henrick K, Berman HM (2005) PDBML: the representation of archival macromolecular structure data in XML. Bioinformatics 21:988–992
7. Henrick K, Feng Z, Bluhm WF et al (2008) Remediation of the protein data bank archive. Nucleic Acids Res 36:D426–D433
8. Velankar S, Dana JM, Jacobsen J et al (2013) SIFTS: structure integration with function, taxonomy and sequences resource. Nucleic Acids Res 41:D483–D489
9. Read RJ, Adams PD, Arendall WB 3rd et al (2011) A new generation of crystallographic validation tools for the protein data bank. Structure 19:1395–1412
10. Montelione GT, Nilges M, Bax A et al (2013) Recommendations of the wwPDB NMR Validation Task Force. Structure 21:1563–1570
11. Henderson R, Sali A, Baker ML et al (2012) Outcome of the first electron microscopy validation task force meeting. Structure 20:205–214
12. Brändén C-I, Jones TA (1990) Between objectivity and subjectivity. Nature 343:687–689
13. Hooft RW, Vriend G, Sander C, Abola EE (1996) Errors in protein structures. Nature 381:272
14. Kleywegt GJ (2000) Validation of protein crystal structures. Acta Crystallogr D Biol Crystallogr 56:249–265
15. Laskowski RA (2009) Structural quality assurance. In: Gu J, Bourne PE (eds) Structural bioinformatics, 2nd edn. Wiley, New Jersey, pp 341–375
16. Brown EN, Ramaswamy S (2007) Quality of protein crystal structures. Acta Crystallogr D Biol Crystallogr 63:941–950
17. Krissinel E, Henrick K (2007) Inference of macromolecular assemblies from crystalline state. J Mol Biol 372:774–797
18. Rose PW, Prlic A, Bi C et al (2015) The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. Nucleic Acids Res 43:D345–D356
19. Finn RD, Tate J, Mistry J et al (2008) The Pfam protein families database. Nucleic Acids Res 36:D281–D288
20. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247:536–540
21. Lovell SC, Davis IW, Arendall WB 3rd et al (2003) Structure validation by Calpha geometry: phi, psi and Cbeta deviation. Proteins 50:437–450
22. Kleywegt GJ, Harris MR, Zou JY, Taylor TC, Wahlby A, Jones TA (2004) The Uppsala Electron-Density Server. Acta Crystallogr D Biol Crystallogr 60:2240–2249
23. Moreland JL, Gramada A, Buzko OV, Zhang Q, Bourne PE (2005) The Molecular Biology Toolkit (MBT): a modular platform for developing molecular visualization applications. BMC Bioinformatics 6:21
24. Stierand K, Maass PC, Rarey M (2006) Molecular complexes at a glance: automated generation of two-dimensional complex diagrams. Bioinformatics 22:1710–1716
25. Goodsell DS, Dutta S, Zardecki C, Voigt M, Berman HM, Burley SK (2015) The RCSB PDB "Molecule of the Month": inspiring a molecular view of biology. PLoS Biol 13, e1002140
26. Gutmanas A, Alhroub Y, Battle GM et al (2014) PDBe: Protein Data Bank in Europe. Nucleic Acids Res 42:D285–D291
27. Krissinel E, Henrick K (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallogr D Biol Crystallogr 60:2256–2268
28. Golovin A, Henrick K (2008) MSDmotif: exploring protein sites and motifs. BMC Bioinformatics 9:312
29. Golovin A, Henrick K (2009) Chemical substructure search in SQL. J Chem Inf Model 49:22–27
30. Reichert J, The SJ, IMB (2002) Jena Image Library of Biological Macromolecules: 2002 update. Nucleic Acids Res 30:253–254
31. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM (1997) CATH: a hierarchic classification of protein domain structures. Structure 5:1093–1108
32. Laskowski RA, Hutchinson EG, Michie AD, Wallace AC, Jones ML, Thornton JM (1997) PDBsum: a web-based database of summaries and analyses of all PDB structures. Trends Biochem Sci 22:488–490
33. de Beer TA, Berka K, Thornton JM, Laskowski RA (2014) PDBsum additions. Nucleic Acids Res 42:D292–D296
34. Laskowski RA, MacArthur MW, Moss DS, Thornton JM (1993) PROCHECK - a program to check the stereochemical quality of protein structures. J Appl Crystallogr 26:283–291
35. Laskowski RA (2007) Enhancing the functional annotation of PDB structures in PDBsum using key figures extracted from the literature. Bioinformatics 23:1824–1827

36. Porter CT, Bartlett GJ, Thornton JM (2004) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. Nucleic Acids Res 32: D129–D133
37. Sigrist CJ, de Castro E, Cerutti L et al (2012) New and continuing developments at PROSITE. Nucleic Acids Res 41:D344–D347
38. Glaser F, Pupko T, Paz I et al (2003) ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. Bioinformatics 19:163–164
39. Wallace AC, Laskowski RA, Thornton JM (1995) LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. Protein Eng 8:127–134
40. Luscombe NM, Laskowski RA, Thornton JM (1997) NUCPLOT: a program to generate schematic diagrams of protein-nucleic acid interactions. Nucleic Acids Res 25:4940–4945
41. Pakseresht N, Alako B, Amid C et al (2014) Assembly information services in the European Nucleotide Archive. Nucleic Acids Res 42: D38–D43
42. Biasini M, Bienert S, Waterhouse A et al (2014) SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Res 42:W252–W258
43. Kiefer F, Arnold K, Kunzli M, Bordoli L, Schwede T (2009) The SWISS-MODEL Repository and associated resources. Nucleic Acids Res 37:D387–D392
44. Pieper U, Webb BM, Dong GQ et al (2014) ModBase, a database of annotated comparative protein structure models and associated resources. Nucleic Acids Res 42:D336–D346
45. Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A (2014) Critical assessment of methods of protein structure prediction (CASP)--round x. Proteins 82(Suppl 2):1–6
46. Marsden RL, Ranea JA, Sillero A et al (2006) Exploiting protein structure data to explore the evolution of protein function and biological complexity. Philos Trans R Soc Lond B Biol Sci 361:425–440
47. Das S, Lee D, Sillitoe I, Dawson NL, Lees JG, Orengo CA (2015) Functional classification of CATH superfamilies: a domain-based approach for protein function annotation. Bioinformatics 31(21):3460–3467
48. Jefferson ER, Walsh TP, Barton GJ (2008) A comparison of SCOP and CATH with respect to domain-domain interactions. Proteins 70:54–62
49. Kolodny R, Petrey D, Honig B (2006) Protein structure comparison: implications for the nature of 'fold space', and structure and function prediction. Curr Opin Struct Biol 16:393–398
50. Prakash A, Bateman A (2015) Domain atrophy creates rare cases of functional partial protein domains. Genome Biol 16:88
51. Orengo CA, Jones DT, Thornton JM (1994) Protein superfamilies and domain superfolds. Nature 372:631–634
52. Novotny M, Madsen D, Kleywegt GJ (2004) Evaluation of protein fold comparison servers. Proteins 54:260–270
53. Carugo O (2006) Rapid methods for comparing protein structures and scanning structure databases. Curr Bioinformatics 1:75–83
54. Joosten RP, Long F, Murshudov GN, Perrakis A (2014) The PDB_REDO server for macromolecular structure model optimization. IUCrJ 1: 213–220
55. Madej T, Lanczycki CJ, Zhang D et al (2014) MMDB and VAST+: tracking structural similarities between macromolecular complexes. Nucleic Acids Res 42:D297–D303
56. OCA, a browser-database for protein structure/function. 1996. (Accessed at http://oca.weizmann.ac.il)
57. Kinjo AR, Suzuki H, Yamashita R et al (2012) Protein Data Bank Japan (PDBj): maintaining a structural data archive and resource description framework format. Nucleic Acids Res 40:D453–D460
58. Bates PA, Kelley LA, MacCallum RM, Sternberg MJ (2001) Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM. Proteins Suppl 5:39–46
59. Nielsen M, Lundegaard C, Lund O, Petersen TN (2010) CPHmodels-3.0--remote homology modeling using structure-guided sequence profiles. Nucleic Acids Res 38:W576–W581
60. Lambert C, Leonard N, De Bolle X, Depiereux E (2002) ESyPred3D: Prediction of proteins 3D structures. Bioinformatics 18:1250–1256
61. Haas J, Roth S, Arnold K, et al (2013) The Protein Model Portal--a comprehensive resource for protein structure and model information. Database (Oxford) 2013;2013:bat031
62. Sillitoe I, Lewis TE, Cuff A et al (2015) CATH: comprehensive structural and functional annotations for genome sequences. Nucleic Acids Res 43:D376–D381
63. Andreeva A, Howorth D, Chothia C, Kulesha E, Murzin AG (2014) SCOP2 prototype: a new approach to protein structure mining. Nucleic Acids Res 42:D310–D314

64. Prlic A, Bliven S, Rose PW et al (2010) Pre-calculated protein structure alignments at the RCSB PDB website. Bioinformatics 26:2983–2985
65. Holm L, Rosenstrom P (2010) Dali server: conservation mapping in 3D. Nucleic Acids Res 38:W545–W549
66. Marti-Renom MA, Pieper U, Madhusudhan MS et al (2007) DBAli tools: mining the protein structure space. Nucleic Acids Res 35: W393–W397
67. Kawabata T (2003) MATRAS: a program for protein 3D structure comparison. Nucleic Acids Res 31:3367–3369
68. Martin AC (2000) The ups and downs of protein topology; rapid comparison of protein structure. Protein Eng 13:829–837
69. Fox NK, Brenner SE, Chandonia JM (2014) SCOPe: Structural Classification of Proteins--extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic Acids Res 42:D304–D309
70. Wang G, Dunbrack RL Jr (2003) PISCES: a protein sequence culling server. Bioinformatics 19:1589–1591
71. Levy ED, Pereira-Leal JB, Chothia C, Teichmann SA (2006) 3D complex: a structural classification of protein complexes. PLoS Comput Biol 2, e155
72. Flores S, Echols N, Milburn D et al (2006) The Database of Macromolecular Motions: new features added at the decade mark. Nucleic Acids Res 34:D296–D301
73. Lomize MA, Lomize AL, Pogozheva ID, Mosberg HI (2006) OPM: orientations of proteins in membranes database. Bioinformatics 22:623–625
74. Lai YL, Chen CC, Hwang JK (2012) pKNOT v. 2: the protein KNOT web server. Nucleic Acids Res 40:W228–W231
75. Kolesov G, Virnau P, Kardar M, Mirny LA (2007) Protein knot server: detection of knots in protein structures. Nucleic Acids Res 35: W425–W428

Springer
http://www.springer.com/978-1-4939-3570-3

Data Mining Techniques for the Life Sciences
Carugo, O.; Eisenhaber, F. (Eds.)
2016, XIII, 552 p. 97 illus., 84 illus. in color., Hardcover
ISBN: 978-1-4939-3570-3
A product of Humana Press