Chapter 2
Structures

2.1 Introduction

The three-dimensional (3D) structure of a protein contains a lot of information on its function, and can be used for devising ways of modifying it (propose mutants, protein design, etc.).

This chapter covers some of the bioinformatics technologies aimed at managing, storing and making computations on protein 3D structures. We will start by getting familiar with the on-line resources where primary and derived information on protein structures is stored, so that we can search and retrieve the available structural information for our protein(s) of interest. Then we will learn the basis on how to manipulate visualize and compare protein 3D structures. Since the experimental structural information is limited to a relatively small number of proteins, we will learn how to predict structural features for raw protein sequences, from low level features such as secondary structure, coiled-coil regions, etc. to complete 3D structures for protein chains. Finally we will take a look at some bioinformatics methods aimed at extracting useful information from these 3D structures (either experimental or predicted).

2.1.1 Storing Protein Structures—The PDB File Format

In the same way protein sequences are stored in text files with particular formats so that they can be interchanged between different programs, there is a widely used format for computationally store protein 3D structures: the "protein data bank" (PDB) format. Almost all programs and web servers working with protein structures are able to handle PDB files. In most cases you will manage PDB files (e.g. for interchanging structural information between diverse servers) without caring about their content or format. Nevertheless, it may be useful to be familiar with their format since some operations with protein structures can be performed by directly editing these files.

F. Pazos, M. Chagoyen, *Practical Protein Bioinformatics,*
DOI 10.1007/978-3-319-12727-9_2

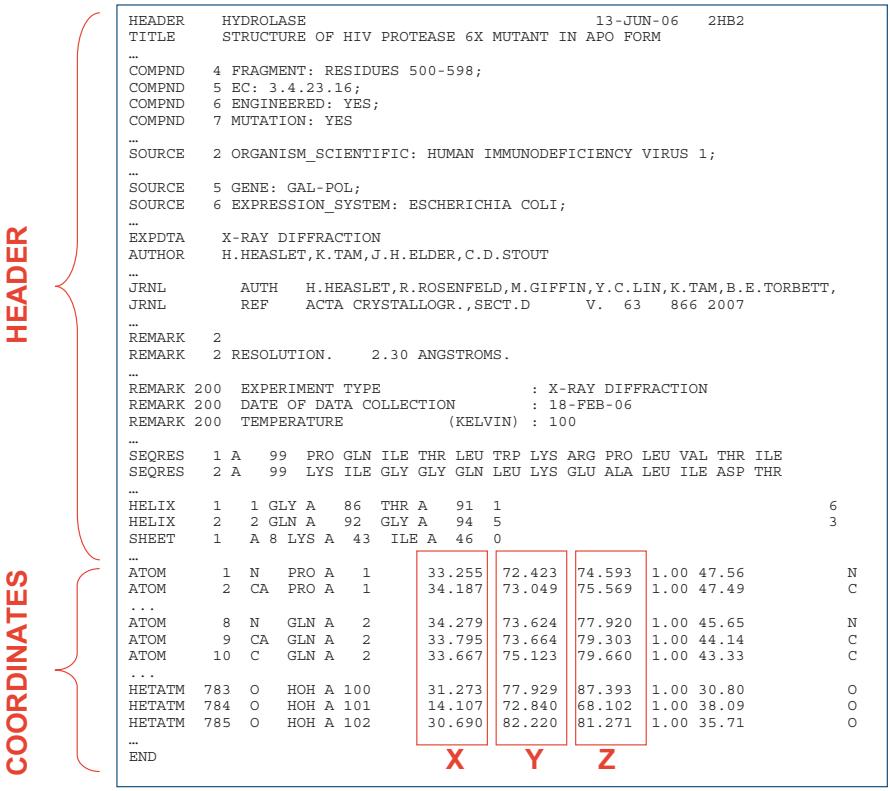

Fig. 2.1 Representative sections of a PDB file for storing information on macromolecular structures

A PDB file is a plain-text file in which the information related to a macromolecular structure involving one or more proteins is stored. Basically, it contains a header section with the "metadata" associated to that structure (protein name, experimental method used for structure determination, …) followed by the Cartesian coordinates (X, Y, and Z) of the atoms (Fig. 2.1).

The header section contains information on the biomolecules whose structures are represented in the file (name, organism, mutations, amino acid sequences, IDs in sequence databases…), the experimental/computational method used for determining/predicting it (e.g. X-ray crystallography, NMR, computational modeling, …) as well as the parameters and experimental details associated to these methodologies. This section also contains the bibliographic reference of the structure determination and information on non-protein molecules present in the structure (cofactors, ions, water molecules …). It may also contain information on the secondary structure content and the binding/functional residues annotated by the generators of the structure.

In the coordinates section (lines starting with "ATOM"), each atom is represented in a line which contain the atom number, its type within the protein (e.g. N: backbone nitrogen; CA: alpha carbon; CB: beta carbon…), the residue type it belongs to and its number within the sequence, the chain (for files containing various protein polymers –complexes, multimers, …-) and the X, Y and Z coordinates in angstrom (Å). The rest of the line contains variable information depending on the method: e.g. B-factor for X-ray structures, confidence figures for predicted models, etc. The "HETATM" lines contain the coordinates of the atoms of non-protein molecules, such as prostetic groups and waters (Fig. 2.1).

If the PDB file contains a number of different alternative structures for the same protein (e.g. alternative computational models or an ensemble of structures compatible with NMR data) these are stored in different "MODEL" sections.

Editing this file it is possible, for example, to extract a chain of interest from a file with multiple chains: simply copying/pasting the ATOM and HETATM lines for that chain to another file. Similarly, by copying/pasting we can extract a representative model from a PDB file representing an ensemble of alternative structures (i.e. NMR). It is also possible to manually remove solvent/water molecules or the atom lines of the residues which are causing problems in some programs.

2.2 Main Protein Structure Databases

There are many on-line databases with primary and derived information on protein three-dimensional (3D) structures.

The original database in which primary information on protein 3D structures is deposited is the ***Protein Data Bank (PDB)*** at the RCSB (Berman et al. 2000).

PDB—Main primary database on protein 3D structures	http://www.rcsb.org http://csbg.cnb.csic.es/PB/T1010	

Apart from hosting the raw data on protein structures, a number of browsing, searching and analysis features were incorporated to the PDB with the time. For a molecular biologist, the main entry points to this massive amount of structural information are the "Search" and "Explore archive" sections of the web interface.

The search form, at the top of the main page, contains a single text entry box for quick searches based on the name of the protein, its sequence, the ligand co-crystallized, etc. Examples are provided when selecting these different search criteria. A more advanced search form is available at "Advanced search". This allows creating complex queries combining an arbitrary number of criteria. For that, select

a search criterion in the "Choose a query type" selector and enter the search value(s) in the corresponding box (for example "structure title"/"hemoglobin"). Then add extra search panels ("+" button) and fill them with additional pairs search criterion/value (e.g. "chemical ID"/"HEM" and "X-ray resolution"/"between 0.0 and 2.0"). The combination of these three search criteria results in the complex query "structures of hemoglobins with resolution 2.0 Å or better and bound to heme groups". Within each panel, the "Result Count" button shows how many entries fulfill that particular search criterion, independently of the others. Once the complete query is constructed, press "Submit query". Close to that button, you have the option of filtering the results by sequence identity, so as to avoid retrieving, for example, many times the same protein crystallized with different cofactors. The results page lists the entries fulfilling all the search criteria together. At the top of the results page, a "filter refinements" panel allows further filtering the results (by organisms, etc.)

A particularly useful search criterion is "sequence (BLAST/FASTA/PSI-BLAST)". This allows looking for proteins of known structure (PDB entries) similar to a sequence of interest. This is the more direct way to check whether our protein has been crystallized, either itself or a close homolog that would allow building a 3D model by homology (Sect. 2.5.1). In the panel associated to that search criterion, paste the sequence of your protein and select the sequence search method as well as the cutoffs of minimum sequence identity or maximum E-value (see Sect. 1.6.2.1). Note that this can also be done in most BLAST servers selecting "PDB" or "RCSB" as the database to search against. The advantage of using this option within the RCSB search page is that the result of the sequence search can be, in a single shot, filtered by additional structural criteria adding more panels as explained above. This allows, for example, looking for homologous of known structure of our protein of interest crystallized with good resolution in a particular model organism.

With the "Explore archive" panel, you can successively narrow the scope of the search click by click until you reach the desired entry(ies). For the example of the hemoglobins above: release date: 2010-today (~38,000 entries) > organism: homo sapiens (~10,000 entries) > experimental method: solution NMR (~700 entries) > enzyme classification: isomerases (11 entries).

In all these cases, you finally end up in a list of entries of interest. In this list, for each entry you can see its PDB ID. This 4-character code (being the 1st character always a digit) is the main identifier of the entry and the standard way of referencing it. There is also a small picture of the structure and some basic information of the entry (title, release date, bibliographic reference, …). Just below the PDB code, there is a link to download the raw PDB file with the whole entry in text format. We will need this file if we want to further use the 3D structure outside the RCSB. Another link below the PDB code ("3D view") opens an interactive 3D viewer (JMol) where the structure can be inspected (rotated, zoomed, …) (Sect. 2.3.1). Finally, clicking the title of the entry, the page with all the information associated to that entry appears.

This information is split in different sections (tabs at the top). Within the "Summary" tab, we find additional images of the structure on the right, including the (eventually different) asymmetric unit of the crystal and the predicted biological

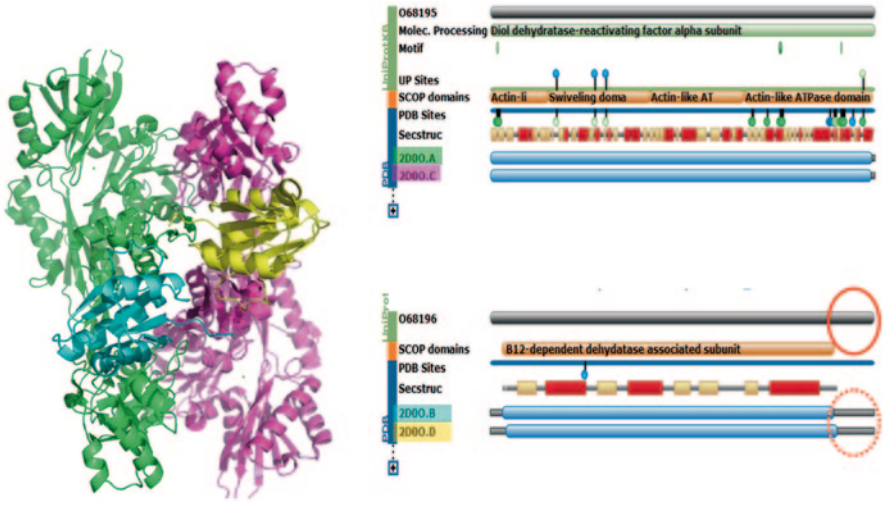

Fig. 2.2 Three-dimensional structure and RCSB molecular description for PDB entry 2D0O. This entry contains 4 protein chains: two dimers of the large and small subunits of the diol dehydratase-reactivating factor. Chains A and C (*green* and *pink*) are the large subunits, and B and D (*blue* and *yellow*) the small ones. Consequently, there are two different proteins within this entry (Uniprot IDs O68195 and O68196-right-). The molecular description panel also shows that the C-terminal part of the small subunits (O68196, chains B and D) has not been crystallized (*red circles*)

assembly. This panel also contains links to visualize/manipulate the structure in a number of structure viewers (Sect. 2.3.1). Another interesting panel of the "Summary" tab is "Molecular description", which contains a linear representation of the protein chains represented by that entry (together with their accessions in Uniprot) indicating which segments are actually in the PDB file (i.e. for which the 3D structure was determined), together with additional features for the proteins (active sites, domains, secondary structure, etc). This provides a quick overview on the global functional and structural characteristics of the proteins, as well as on the parts of the protein sequence which are actually crystallized (e.g. the structure could have been determined for a segment of the protein only) (Fig. 2.2).

The "Sequence" tab displays more detailed information on the sequence of the protein chains, together with their secondary structures and the residues involved in binding ligands. Finally, the "Links" tab contains links to the pages for that entry in other structure-related databases, mainly secondary databases based on PDB.

One of such PDB-derived databases is the ***PDBsum*** at the EBI (Laskowski et al. 2005). The original goal of PDBsum is to "enrich" the raw structural information of the PDB adding representations and data extracted and/or calculated from it. Although, as commented above, the current version of PDB also includes a lot of derived data, there are still a number interesting and unique features in PDBsum.

PDBsum—Derived data on protein 3D structures	http://www.ebi.ac.uk/pdbsum/ http://csbg.cnb.csic.es/PB/T1020	

PDBsum's top page allows to search by the 4-character PDB code, and by text using other information associated to the PDB entries and the protein chains within them (protein name, author, IDs in different sequence databases, …). It is also possible to look for entries similar to a given sequence provided by the user, as explained earlier for PDB.

The main page for a given entry shows on the left a picture of the structure and a summary of the molecules within it (protein chains, DNA/RNA chains, heteroatoms –ligands-, metals, waters, etc.). Three orthogonal views of the structure are available with the "eye" icons, and a link to interactively inspect that structure in a JMol viewer is also provided (see Sect. 2.3.1). On the middle column, you find a summary of the fields within the header of the PDB file (title, molecule name, author, …) and a "Links" section with hyperlinks to the corresponding pages for that PDB code in other structural databases). Here you also find a linear representation of the protein chains within that entry highlighting their coverage respect to the whole-length protein deposited in the sequence database, the annotated domains, functional sites, secondary structure, etc., similar to the "molecular description panel" of the RCSB commented above. On the right, we find the Ramachandran plot of the structure, which is a link to a complete report of Procheck (Laskowski et al. 1996), a program for evaluating the quality of protein structures. This top page also shows the abstract of the publication describing the determination of the structure and selected figures from that article, as well as a list of other papers citing it.

The "Protein" tab contains detailed information, at the residue level, of the protein chains within the entry (conservation, secondary structure, etc.). Here, the protein chains are also split in domains according with the CATH database (Sect. 2.2.1) and representations of the topology of these domains are shown on the right.

The "DNA/RNA", "Ligand" and "Metal" tabs contain diagrams showing the interactions between these biological entities and the protein chains in the structure (type of interaction, protein residues involved, …). The diagrams are generated with the LIGPLOT software (Wallace et al. 1995). For multi-chain PDB entries, the "Prot-prot" tab contains information on the interaction surfaces (interfaces): size, residues involved, … Similarly, the "Clefts" and "Tunnels" tabs contain the elements of this kind detected in the structure. These elements are related to binding and functional sites in many cases and, in combination with other evidences (clustering of residue conservation, etc.) can help in the prediction of these important regions.

2.2.1 Classifications of Structural Domains

Both PDB and PDBsum store protein structures not framed in any classification schema. Some PDB-derived databases try to classify the PDB entries in a hierarchical schema based on their structures, sequences, and inferred evolutionary relationships. Knowing the position of our protein of interest in these classifications might provide important information on its evolution, function and relatives.

Since domains are the structural, functional and evolutionary units of proteins, the protein structure classification schemas we are going to discuss below have the domain as the basic unit, and not the protein chain or the PDB entry (which can contain multiple chains eventually of different proteins). For example, the PDB entry 2d0o (Fig. 2.2) contains 4 protein chains: two hetero-dimers of the large and small subunits of the diol dehydratase reactivating factor. The small subunits (two chains within this entry) are mono-domain chains, while the large subunits (the other two chains) are chains with two different domains. Consequently, there are three different protein domains within this PDB entry, and indeed it is associated to three entries in the structural classification databases.

One of such databases is the ***Structural Classification of Proteins (SCOP)*** at the MRC (Andreeva et al. 2004).

SCOP—Hierarchical classification of protein domains of known 3D structure	http://scop.mrc-lmb.cam.ac.uk/scop/index.html http://csbg.cnb.csic.es/PB/T1030	

SCOP classifies protein domains according with a hierarchical schema in which at the top level we find the "structural classes". A given structural class is divided into "folds", which in turn contain "superfamilies". Superfamilies can be split into "families" and, finally, these contain the "domains" (Fig. 2.3).

Structural classes group domains according with their global composition of secondary structure elements (α-helix, β-strand, turns, etc.) Consequently, a class contains all the domains comprising α-helices only, other those comprising a mixture of α-helices and β-strands, etc. Some classes may also reflect broad aspects of the arrangement of the secondary structure elements. Classes are clearly artificial since, in general, they do not respond to any evolutionary or functional criteria. Moreover, in an attempt to comprehensively include all proteins available in PDB, this level comprises even more artificial classes defined based on methodological aspects, such as "low resolution structures", "small peptides", etc.

A structural class is subdivided into "folds". A fold groups those domains with the same content and 3D arrangement of secondary structure elements. For example, within the "all-α" structural class we find many different folds depending on the number and relative orientations of the α-helices. Folds are not homogeneous

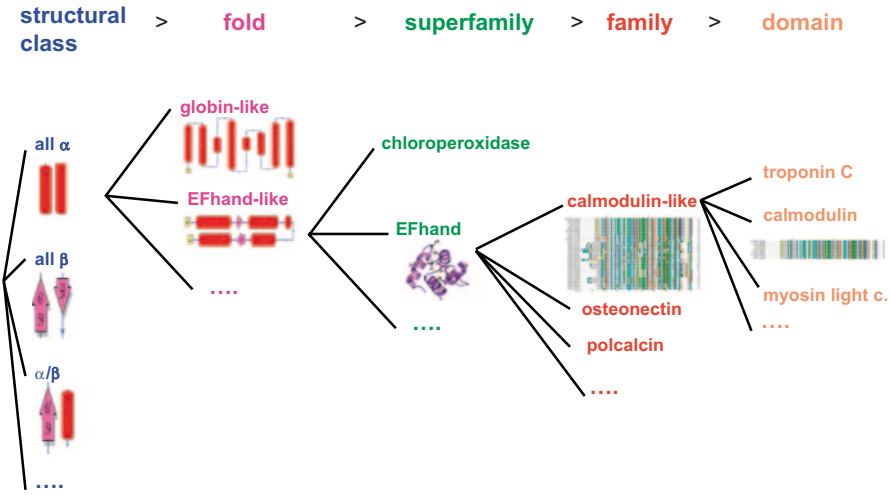

Fig. 2.3 SCOP hierarchical schema for classifying protein domains

in evolutionary or functional terms, in general. A given fold may comprise proteins with very different evolutionary origins and functions. However, it is at this level where the classification starts having some biological rationale since folds represent stable 3D layouts of protein chains repeatedly used by evolution due to their stability and functional advantages.

Within a given fold, superfamilies group homologous domains, that is, domains arisen from a common ancestor. The assignment of homology may be based on a clear sequence identity between the two domains or on subtler evidences, such as short sequence and structural motifs, and/or expert knowledge. For example, the "spectrin repeat-like" fold (all-α class) contains 16 different superfamilies, which are evolutionary unrelated even if they all have the same global 3D structure (number and arrangement of the α-helices). Domains within the same superfamily also share functional aspects, retained from the ancestor they all come from. We could say that the superfamily is the most interesting level from a practical point of view, since functional similarities (even if subtle) start to be forged here. Regarding structural similarities, even if they start at the fold level, in practical terms predicting fold in absence of any sequence signal is very difficult (Sect. 2.5). Consequently, the superfamily provides the best context for inferring structural and functional clues about our protein of interest.

Going deeper in the hierarchy, families group proteins with a clear sequence relationship. The functional similarity between members of a family is higher than between members of a superfamily. While both, families and superfamilies, reflect homology relationships (common ancestry), usually these are referred to as "close homology" and "remote homology" respectively. Finally, a given family contains the individual protein domains (the domain in different organisms (orthologs), mutants and other variations of the same sequence in PDB, etc.), and can eventually be be further "subdivided" into "subfamilies".

SCOP's web interface is very simple and the two main entry points to this resource are the "Keyword search" and "Enter SCOP at the top of the hierarchy" options, which are in the "Access methods" section of the home page.

Entering SCOP at the top of its hierarchy shows a list with the structural classes (first level in SCOP's hierarchy). Clicking a class expands it in the folds it contains, which in turn can be expanded to superfamilies, and so on up to the individual domains. For all the levels there are links to interactively visualize the corresponding domain (or a representative of the fold, superfamily, …). For the superfamilies, there are links to the corresponding entry in the Superfamily resources (section 1.9).

For users not familiar with SCOP's internal IDs and keywords, the "Search" facility is mainly limited to PDB IDs. As commented above, searching for a given PDB ID might result in several SCOP entries depending on the number of individual domains represented by that particular structure. If we want to locate a sequence of interest (or a close relative) in SCOP, it is better to look for it in PDB or other resources which provide much better search features (including by sequence-similarity) and then follow the outgoing links to SCOP.

The ***CATH*** resource at the UCL (Pearl et al. 2005) is also intended to classifying all known protein domains in a hierarchical structure. There are some conceptual and implementation differences with SCOP and, consequently, these two resources complement each other.

CATH—Hierarchical classification and annotation of protein domains	http://www.cathdb.info/ http://csbg.cnb.csic.es/PB/T1040	

CATH's hierarchy is slightly different to SCOP's, and its levels are "class", "architechture", "topology" and "homologous superfamily", followed by a number of "subfamily" levels defined based on arbitrary thresholds of sequence identity. The "topology" and "homologous superfamily" levels are equivalent to the "fold" and "superfamily" levels of SCOP. Another difference is that CATH has a strong emphasis in annotating raw genomic sequences based on matching against the profiles of its superfamilies (implicitly annotating also at the topology and upper levels). Consequently, CATH extends beyond proteins of known structure and includes all raw sequences that can be matched against its superfamilies. SCOP relies on external resources (SUPERFAMILY, Sect. 1.9) for that. Another characteristic of CATH is the integration of functional information from other resources (EC number, GeneOntology annotations, …) into its entries.

As in SCOP, the main entry points to CATH are the "Browse" and "Search" options, at the top of its main page. The Browse option allows navigating CATH hierarchy by expanding/collapsing nodes within the different levels. During this navigation, a panel on the left displays the information of the expanded node. The

"Search" form allows looking for CATH entries based on a number of different IDs (including CATH internal IDs, PDB IDs, etc.) as well as on keywords. There is also the possibility of directly searching by sequence similarity. Another interesting possibility is to search for structural similarity ("Search by PDB structure"). The user uploads a protein structure in PDB format and the system looks for CATH entries with similar 3D structure (see Sect. 2.3.2.2).

2.3 Structure Manipulation, Visualization and Comparison

In the previous section, we have listed some resources which can be used to look for the available structural information of your protein of interest and its relatives, as well as to put it in the context of a protein structural/evolutionary classification. In that section we also tangentially touched the issues of visualizing, manipulating and, to a lesser extent, comparing protein structures. All these will be expanded in this section.

Note that most tools commented in this section, at least those that admit a generic structure in PDB format as input, can be used with experimentally determined structures as well as structural models (predicted structures, Sect. 2.5).

2.3.1 Structure Manipulation and Visualization

Almost all bioinformatics studies including protein 3D structures involve, at some point, manipulating and visualizing them.

The best software for visualizing and manipulating protein structures comprises stand-alone programs which, while not difficult to use, do not run within a web browser and have to be locally installed in an operative system dependent manner. Consequently, they do not fall within the scope of this book. Nevertheless, due to the large difference in capabilities with the browser-based solutions we are going to comment here, we recommend readers particularly interested in this subject to explore these solutions, for example PyMol (http://www.pymol.org/).

The most widely used molecular visualizer designed to run within a web browser is ***JMol***. JMol runs as a Java applet embedded in a web page together with other elements. Normally, JMol applets show up "preloaded" with a 3D structure, for example in the web page of the corresponding entry for that structure in the resources commented in Sect. 2.2. But it can also be used to load files with structures provided by the user (e.g. a predicted structure). Additionally, JMol is highly customizable (size, contents of the menus, etc.) and can be connected with other elements in the web page, which makes different JMol applets to look slightly different, depending on the web page they are embedded and the purpose of the molecular visualization in that particular page. Although, the web address for JMol provided below is that

http://www.springer.com/978-3-319-12726-2

Practical Protein Bioinformatics

Pazos, F.; Chagoyen, M.

2015, VIII, 106 p. 40 illus. in color., Hardcover
ISBN: 978-3-319-12726-2