PDB

The Protein Data Bank

A CRITICAL GUIDE

Version: 1 August 2018

PDB

Overview

This Critical Guide provides a brief outline of the Protein Data Bank – the PDB – the world's primary repository of biological macromolecular structures. The rationale for creating the resource and the kinds of information it provides are discussed, and issues relating to its evolution and growth are explored.

Teaching Goals & Learning Outcomes

This Guide introduces the principal features of the PDB, the nature (and quality) of its contents and how these may be interrogated. On reading this Guide, you will be able to:

- explain some of the ways in which knowledge of protein structures is useful;
- identify the constituent databases of the wwPDB;
- explain key features of the RCSB PDB in terms of its data distribution, growth and redundancy statistics;
- search the PDB using simple and advanced keywords and full sequences, and analyse differences between them; and
- explain various structural quality criteria, and infer the quality of individual PDB entries.

1 Introduction

The Protein Data Bank (PDB)[1] can probably be considered the first bio-database. It was created in October 1971, long before computers were commonplace tools in the workplace, before even the term 'bioinformatics' had came into common usage and the field was recognised as a discipline in its own right. The PDB was modelled on the **Cambridge Structural Database** (CSD)[2], which had been created in the mid 1960s, under the direction of **Olga Kennard**, to collect **small-molecule** structures derived using **X-ray crystallography**. The idea was to start collating and storing the 3D coordinate data for proteins, which were beginning to present challenging new horizons across the biochemistry and biophysics communities. The vision was to be able to exploit the collected protein structures to discover new knowledge[3]. Although this may now seem a rather trivial or self-evident goal, the idea of creating a 'bank' of protein structures, similar to the CSD, was actually quite radical at a time when computers were extremely rare. Interestingly, Kennard also envisaged the PDB as a paradigm for the European nucleotide sequence data library – a resource that took 10 further years to materialise[4].

Possibly because the concept of electronic databases was so far ahead of its time, buy-in from crystallographers was slow. Launched with 7 structures[5], the PDB still only housed 9 structures when it became fully operational two years later[6], despite many more structures having been determined and published. By 1977, the contents had grown to 47 macromolecules, including the first **transfer RNA** (tRNA)[7]. The inclusion of non-protein data sparked a serious debate about whether the *Protein Data Bank* was still a relevant name; but the brand was well known by this time, and the name stuck. The *Protein* Data Bank thus continues to house nucleic acid and protein-nucleic acid complexes, which now number in the thousands.

But the 1977 holdings pointed to a much more important concern. Owing to the nature of structural biology research, the structure of a given protein is often solved repeatedly in a variety of different experimental contexts. Consequently, the resource was accreting a significant amount of redundancy (estimated to be ~7-fold by 1992[8,9]), which inevitably skewed the reported growth statistics.

Despite the challenges, however, the PDB has grown to become the world's principal archive of biomolecular structures. This Guide introduces the PDB, placing particular emphasis on the nature of the information it provides, outlining how the resource is maintained, and how it may be interrogated.

2 About this Guide

The following sections review some of the main features of the PDB. The Guide isn't a complete tour of the Web interface, as front-ends frequently change; rather, it gives a high-level content overview. Exercises are provided to help understand how to navigate and search the information stored in the PDB, and how to discover the quality of its entries. Throughout the text, key terms – rendered in **bold** type – are defined in boxes. Additional information is provided in supplementary boxes.

KEY TERMS

Cambridge Structural Database (CSD): archive of small molecule crystal structures compiled & maintained by the Cambridge Crystallographic Data Centre

Olga Kennard: crystallographer, Director of the Cambridge Crystallographic Data Centre from 1965-97, Fellow of the Royal Society & OBE

Transfer RNA (tRNA): small adaptor molecule that mediates the synthesis of proteins, carrying specific amino acids to the designated codon specified by an mRNA template

Small molecule: in biochemistry, a low-molecular-weight organic compound (not a polymer); in pharmacology, a molecule that binds with high affinity to a biopolymer, altering its activity or function

X-ray crystallography: technique for deducing atomic arrangements in crystals using X-ray diffraction; routinely used to determine the structures of small molecules, proteins & nucleic acids

3 What is the PDB?

The PDB (www.rcsb.org) is the central repository of biomolecular structure data, maintained since 1999 by the **Research Collaboratory for Structural Biology** in the USA. It is a member of the worldwide Protein Data Dank (wwPDB), an umbrella organisation established to harmonise the activities of structure databases around the world, creating a single, global, public archive of macromolecular structure data[10,11]. The wwPDB brings together four main resources:

i) *RSCB PDB*[12]: derived from the original PDB established at the **Brookhaven National Laboratory** in the USA, and now the principal curator of PDB data;
ii) *PDBe*[13]: PDB Europe, the European resource for collecting, organising and disseminating biological macromolecular structural data, maintained at the **European Bioinformatics Institute** (EBI) in the UK;
iii) *PDBj*[14]: PDB Japan, the central Japanese archive of macromolecular structures, maintained at Osaka University, with support from the Japan Science and Technology Agency's **National Bioscience Database Centre**; and
iv) *BMRB*[15]: the Biological Magnetic Resonance Data Bank, a data repository for the structures of proteins, peptides, nucleic acids and other biomolecules determined by **NMR spectroscopy**, maintained at the University of Wisconsin, USA.

The RCSB remains the 'archive keeper', having sole write-access to the PDB, controlling its contents, and distributing new PDB **identifiers** (IDs) to all deposition sites.

3.1 The PDB holdings

Structural studies are extremely valuable, not just for the evolutionary insights they provide about the physical similarities shared by some proteins, but also for offering a basis on which to understand the mechanistic processes via which proteins effect their functions. In an ideal world, 3D structural coordinates would be available for every known protein sequence; however, there are almost 1000-fold more sequences in **UniProtKB**[16] than there are structures available in the PDB, reflecting a significant sequence-structure deficit.

As noted earlier, the PDB initially grew very slowly: from its original 7 entries in 1971, it had amassed only 13 structures by 1976. The growth trajectories for protein structures during the decades 1976-1985, 1999-2008 and 2009-2018 are noted in Table 1. The figures suggest that the resource saw a 10-fold expansion in its first 10 years, and an overall >10,000-fold increase from 1976 to 2018: by 1 August 2018, the PDB held 3D coordinate data for 132,361 proteins.

Table 1 ***Growth of PDB protein entries from 1976-1985, 1999-2008 & 2009 to 1 August 2018.*** *The number of entries for each year is shown.*

Year	# of Entries	Year	# of Entries	Year	# of Entries
1985	173	2008	50,427	2018	132,361
1984	155	2007	43,906	2017	126,621
1983	134	2006	37,136	2016	116,287
1982	99	2005	31,120	2015	106,278
1981	71	2004	26,116	2014	97,706
1980	57	2003	21,311	2013	88,969
1979	50	2002	17,484	2012	80,192
1978	40	2001	14,722	2011	72,027
1977	36	2000	12,133	2010	64,586
1976	13	1999	9,743	2009	57,325

Interestingly, the statistics shown in Table 1 quote 36 entries in 1977; however, in their paper at the time, Bernstein *et al.* listed 47 macromolecular structures contained in 77 atomic coordinate entries[7]. Notwithstanding the discrepancy, this pointed to a significant amount of redundancy.

3.2 Redundancy in the PDB

As the PDB took on the role as the single worldwide macromolecular structure archive, its growing levels of redundancy caused continued concern. In 1994, an analysis showed that 91% of 1,107 newly deposited structures had identical or highly similar sequences to existing entries, only 9% had no obvious **homologue** in the PDB, and only around a third of those represented new folds[17].

Partly to bolster the number of available structures, and partly to temper the redundancy issues, several pilot studies were run in 1999 to determine the feasibility of a targeted, large-scale, high-throughput structure-determination programme (modelled on the **Human Genome Project**)[17]. The results were sufficiently compelling that a high-throughput X-ray crystallography project was launched the following year, an ambitious, high-profile venture known as the **Protein Structure Initiative** (PSI). The ambition of the PSI was to produce 10,000 new structures over the next ten years, ultimately providing structures, or models, for proteins in all completed genomes. Another motivating principle was the belief that knowledge of all these new structures would divulge their molecular functions.

The reality was rather different: 17 years after its launch, the closing tally of distinct PSI structures was 5,472, and the functions of around a third of those remained unknown – indeed, researchers were invited to become '*functional sleuths*' in a community-wide effort to help characterise them (see section 4).

KEY TERMS

Brookhaven National Laboratory: a national lab based in New York, owned by the US Department of Energy, & former home of the PDB
European Bioinformatics Institute: the EMBL hub dedicated to the provision of bioinformatics services across Europe, based at Hinxton
Homologue: a molecular sequence (or structure) whose level of similarity to another sequence (or structure) suggests shared ancestry; significant levels of similarity provide evidence that sequences (or structures) are related by divergent evolution from a common ancestor
Human Genome Project: international project to map & sequence the human genome, managed by the US Department of Energy & the NIH
Identifier (ID): a unique code used to identify a specific entry in a particular database; typically, ID codes are designed to be more comprehensible to users than their corresponding accession numbers
National Bioscience Database Centre: a national centre based in Tokyo, that integrates life science-related databases worldwide, to optimise the value of scientific data
NMR spectroscopy: a technique for observing local magnetic fields around atomic nuclei; often used to identify proteins & other complex molecules, & provide details of their structures, dynamics, *etc.*
Protein Structure Initiative (PSI): a major federal, university & industry project, launched in 2000, exploiting high-throughput methods to reduce the cost & decrease the time taken to deduce protein structures
Research Collaboratory for Structural Biology (RCSB): a consortium of Rutgers, State University of New Jersey; the San Diego Supercomputer Center, University of California; & Center for Advanced Research in Biotechnology of the National Institute of Standards & Technology
UniProtKB: UniProt Knowledgebase, a protein sequence database comprising UniProtKB/Swiss-Prot & UniProtKB/TrEMBL

Redundancy by sequence similarity

An idea of just how much redundancy is sequestered in the PDB can be gained by clustering its contents according to different levels of sequence identity, using the popular pairwise sequence comparison algorithm, **BLAST**[18]. Table 2 shows the result of clustering the 1 August 2018 holdings. At the extremes, clustering at 100% identity suggests a near 50% redundancy; at 30% identity, it appears that there remain only ~28,000 sequences, a little more than a fifth of the total. These subsets may be searched directly via the query interface, allowing more focused, efficient searches, depending on the level of sequence identity chosen. Nevertheless, caution is advised when interpreting the results, because they are returned on a structure basis, while sequence similarity is defined here on a chain basis – and many PDB structures contain multiple chains.

Table 2 ***PDB's sequence-based redundancy, 1 August 2018, calculated using BLASTClust from the standalone BLAST package.*** *The level of sequence identity & number of non-redundant sequences are shown.*

% Sequence identity	# of Non-redundant sequences
100	68,054
95	55,014
90	52,114
70	45,517
50	38,577
40	33,730
30	28,195

The picture is further complicated because, even at the level of 30% sequence identity, many of the proteins in this subset will have the same **fold**, because structures are more conserved than their underlying sequences. To gain a better understanding of structural similarity within the PDB, it is necessary to analyse its contents using more sophisticated **fold-classification databases** as benchmarks.

Redundancy by structure similarity

A rather different perspective on the amount of redundancy in the PDB is gained by clustering its contents according to different fold classes, for example using the fold-classification database, **CATH**[19]. Figure 1 shows the result of clustering the 1 August 2018 holdings. This benchmark suggests that the number of unique folds in the PDB is 1,375, and has remained at around this level since 2009 – only 16 new folds have been observed in the PDB in the last decade; none have entered the database since 2012.

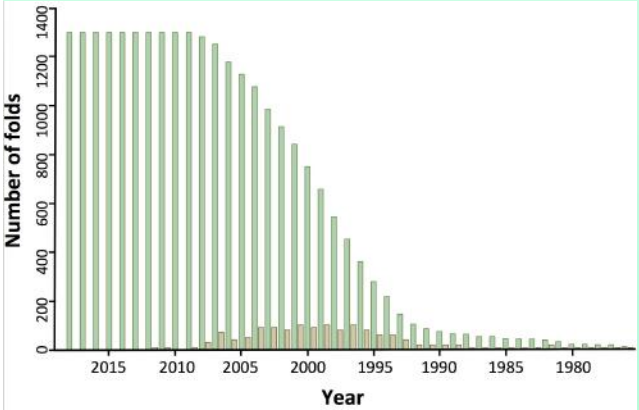

Figure 1 ***PDB's fold-based redundancy, 1 August 2018, calculated using CATH.*** *The total number of unique folds in each year (green), and the yearly addition of new folds (gold) are shown.*

This is interesting from the perspective of the PSI, whose vision included tackling PDB's redundancy issues by targeting novel structures. Despite this ambition, however, as Figure 1 highlights, the number of unique folds entering the database each year did not increase during the decade following the launch of the PSI, and has remained at zero for the last six years.

4 The Structural Biology Knowledgebase

As mentioned earlier, the PSI was a major project launched in 2000 to reduce the cost, and hasten the pace, of protein structure determination. The project was distributed across several Structural Genomics Centres, and a Structural Biology Knowledgebase (SBKB) was developed[20] to coordinate and integrate the data accumulating from the participating organisations. The SBKB permitted BLAST searches with protein sequences, returning similar structures from the PDB, together with theoretical models from the Protein Model Portal[21] and information from more than 150 related databases.

The PSI formally ended in 2015, but continued to deliver results until July 2017; at its close, the final PSI summary reported the deposition of almost 7,000 coordinate sets, of which ~5,500 represented distinct structures, together with more than 20 million **homology models**. With the completion of the PSI, most of the services of the SBKB were terminated, including public access to the portal itself, which was replaced by a simple summary page (sbkb.org).

However, the functionality of the original site can still be accessed via an instance of the portal retained for internal use (kb-new-dev.sbkb.org). Here, it is still possible to search the SBKB, and to accept the challenge of becoming a *functional sleuth*, to help characterise structures whose functions are unknown, or for which only minimal information exists (kb-new-dev.sbkb.org/functionalsleuth).

EXERCISES

1 Using the generic '*Search*' box at the top of the internal PSI SBKB home page, select '*by text*', type the keyword 'rhodopsin' & click on '*Go*'. Make a note of the information returned.

2 Return to the '*Search*' box, select '*by pdb id*', type '1f88' & click on '*Go*'. Note the information returned.

3 Return to the '*Search*' box, select '*by uniprot ac*', type 'P02700' (use AC #s not IDs) & click on '*Go*'. Note the information returned.

4 Return to the '*Search*' box, select '*by sequence*', paste in the sequence (www.uniprot.org/uniprot/P02700.fasta) of UniProtKB entry 'OPSD_SHEEP', delete the FastA header & click on 'Go'. Were the results from searches 1-4 the same? What are the main differences? Which search was the most selective? Which returned the most structures? Which was the least focused? Why might this be so?

KEY TERMS

BLAST: Best Local Alignment Search Tool, a program for searching nucleotide or protein sequence databases with a query sequence

CATH: a resource that classifies protein folds according to their Class, Architecture, Topology & Homology

Fold: the spatial arrangement of protein secondary structures; often used as a basis to classify the tertiary structures of proteins

Fold-classification database: a database that classifies protein structures (including domains within tertiary folds) according to architectural, topological &/or evolutionary similarities

Homology model: a model of a protein derived by mapping its sequence to the 3D structure of a homologous 'target' protein

5 Interrogating the PDB

There are several ways to explore the information stored in the PDB, depending on whether your interest is in discovering whether the structure of a particular protein or structures related to your protein of interest are known, or whether you're interested in proteins that bind a particular ligand or that are the targets of particular drugs. In the sections that follow, we examine the principal ways of interrogating the database online (the Guide does not provide a comprehensive breakdown of the numerous other features of the Web interface). Important to note is the generic *Search* box and set of pull-down menus at the top of the PDB home page, which give access to a range of search options.

5.1 PDB keyword search

The simplest way to interrogate the database is to supply a keyword to the generic *Search* box at the top of the home page: this could be the name of a protein (*e.g.*, **rhodopsin**), a ligand (*e.g.*, **retinal**), or author associated with a given structure (*e.g.*, **Perutz**); or it could be a UniProtKB accession number (*e.g.*, P02699) or PDB ID of a specific protein (*e.g.*, 1f88). An *Advanced Search* link beneath the input box gives many other options (too many to list here), amongst which is the ability to perform a sequence search (see Section 5.2).

When performing keyword searches, results are usually returned with a range of filters to help refine the output: *e.g.*, results of a text search for rhodopsin could be filtered by focusing on those structures relating to a given organism, those determined by NMR, just those at high resolution (see supplementary box, Section 5.3), *etc*.

Depending on the nature of the keyword search performed, outputs may also be filtered by means of a series of tabs provided at the top of each results page. For example, for text searches with terms like rhodopsin or retinal, the full set of returned structures may be narrowed down to view only those with bound ligands, to view those that have not yet been released, to focus on their literature citations, and so on. When searching for a specific structure using its PDB ID, a different set of filters is provided: here, a *Structure Summary* gives an overview of the results, and the associated tabs allow further drilling down: *e.g.*, to view the relevant experimental data; to find similar sequences or structures; to find related information in the fold-classification or family databases; or to generate interactive 3D views – these may be customised to highlight different features of the molecule (with or without water molecules, ligands, ions, *etc*.) using different rendering styles (see **Figure 2**).

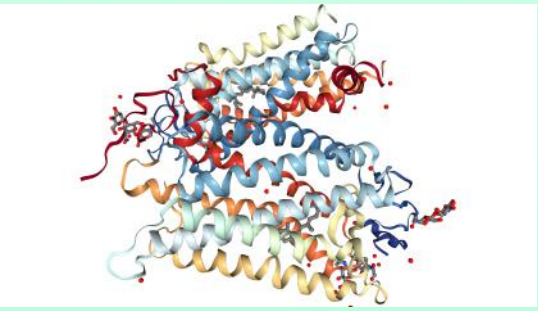

Figure 2 ***Static '3D' view of 1f88.*** *A rainbow cartoon of the structure is depicted, along with the associated ligands & water molecules.*

An alternative way to access search options is via the menu bar at the top of the home page, using the pull-down *Search* menu: this allows, for example, focused searches for specific ligands, drugs (*e.g.*, pilocarpine) and their targets; it provides access to new and unreleased entries; and, again, offers sequence search options. It is important to note that results generated by these more focused searches will differ from those produced by keyword queries in the generic search box. It is beyond the scope of this Guide to detail how to use each of the pull-down menus, but it might be instructive, say, to use the *Ligands* option of the *Search* pull-down menu and provide the name *retinal*, or the *Drugs & Drug Targets* option and provide the name *pilocarpine*, and then compare the outputs with those generated by the equivalent generic keyword searches.

5.2 PDB sequence search

As already mentioned, beyond keyword searches, it is also possible to search the PDB with a query sequence. The input form for sequence searches may be accessed either via the *Advanced Search* link beneath the generic *Search* box at the top of the home page, or via the pull-down *Search* menu in the home-page menu bar (this latter route is the simplest option). Both options perform default BLAST searches; however, the *Advanced Search* allows the default parameters to be customised, including changing the **E-value** cutoff, setting a sequence identity cut-off, masking **low-complexity regions** and changing the algorithm to **PSI-BLAST**.

EXERCISES

1 Using the generic '*Search*' box at the top of the RCSB PDB home page, type the keyword 'rhodopsin' & click on '*Go*'. Make a note of the information returned.

2 Return to the '*Search*' box, type '1f88' & click on '*Go*'. Note the information returned.

3 Return to the home page & click on the '*Search*' pull-down menu in the top menu bar; select the '*Sequences*' option. In the '*Paste Sequence*' box, paste the sequence of UniProtKB entry '*OPSD_SHEEP*' (www.uniprot.org/uniprot/P02700.fasta) & click on '*Run Sequence Search*'. Note the information returned.

4 What are the main differences between the results returned in searches 1-3? Which search was the most selective? Which returned the most structures? Which was the least focused search & why?

5 How do the results compare with the equivalent searches of the SBKB? Why might the results differ?

KEY TERMS

E-value: the number of matches with scores greater than or equal to that of a retrieved match expected to occur by chance in a database of the same size & composition, with the same scoring system

Low-complexity region: part of a protein sequence that has biased composition; generally, these are repetitive regions, or regions enriched in a particular amino acid or group of amino acids

Opsin: light-sensitive visual pigment of cone photoreceptor cells, found in the retinas of most vertebrates; it mediates daylight colour vision

Perutz: a molecular biologist whose pioneering work involved determining the first structures of proteins at atomic resolution

PSI-BLAST: Position-Specific Iterated-BLAST, a version of BLAST in which results from successive database searches contribute to the scoring matrix, increasing its diagnostic power with each iteration

Retinal: or retinaldehyde (vitamin A aldehyde), a poly-unsaturated organic compound that binds to **opsins**, providing the chemical basis for vision in animals

Rhodopsin: light-sensitive visual pigment of rod photoreceptors, found in the retinas of most vertebrates; an achromatic receptor, it mediates vision in dim light

5.3 PDB statistics search

Arguably, one of the most informative ways to interrogate the PDB is to review its data distribution, growth and redundancy statistics. These may be accessed directly from both the *Search* and *Analyze* pull-down menus in the menu bar at the top of the home page.

The distribution of PDB data may be broken down in many different ways: *e.g.*, options here include analysis of the current holdings in terms of i) experimental method and molecular type, ii) enzyme classification, iii) source organism, iv) resolution, v) residue count, vi) contributing Structural Genomics Centres, amongst others.

The data distribution by experimental method and molecular type for the 1 August 2018 holdings is shown in **Table 3**. From the table, it is evident that protein structures dominate PDB's contents, those solved by means of X-ray crystallography being the most abundant.

Table 3 ***Data distribution by experimental method & molecular type.*** *The number of protein, nucleic acid, protein-nucleic acid or other structures is given for each method for the 1 August 2018 holdings.*

Method	Proteins	Nucleic acids	Protein/NA complexes	Other	Total
X-ray	119,545	1,945	6,126	10	**127,626**
NMR	10,793	1,252	250	8	**12,303**
Electron microscopy	1,666	31	589	0	**2,286**
Other	242	4	6	13	**265**
Multi-method	115	4	2	1	**122**
Total	**132,361**	**3,236**	**6,973**	**32**	**142,602**

A sense of the quality of the X-ray and electron microscopy structures can be gained by referring to the distribution of PDB data by resolution, shown in **Figure 3**. Most structures are high-resolution, although many are likely to include small errors: at ~2Å, water molecules and small ligands are discernible; beyond ~2.5Å, surface loops and side-chains are more error-prone; beyond 4Å, resolution is at the level of secondary structures, rather than individual atoms.

Resolution of protein structures

A broad measure of the quality of a protein structure is its 'resolution'. For any imaging system, resolution refers to the extent to which closely adjacent objects can be distinguished as separate items; it hence gives an indication of how much detail may be discerned in the entity being imaged. In the context of imaging protein structures, resolution is determined by the extent to which **electron densities** in a **diffraction pattern** can be distinguished as isolated 'blobs' (atoms), or merge together to appear more like 'tubes' of interconnected atom densities.

Resolution (Å)	Likely structural quality
0.5-1.5	Structures usually have few errors; individual atoms can be resolved
1.5-2.0	Folds seldom incorrect (even in surface loop regions), but many small errors likely
2.0-2.5	Folds usually correct, but many small errors likely; small ligands & water molecules may be seen
2.5-3.0	Folds likely to be correct, but may include errors in surface loops & **side-chains**
3.0-4.0	Folds may be correct, but errors highly likely
>4.0	**Secondary structures** can be determined, but not individual coordinates

The table shows the quality to be expected in structures determined by diffraction techniques, moving from high to lower atomic resolution.

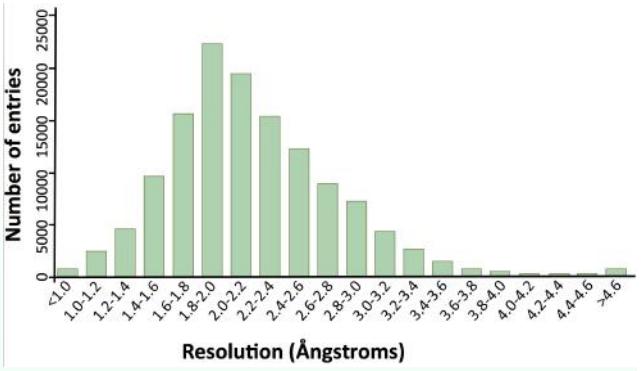

Figure 3 ***PDB data distribution by resolution, 1 August 2018.*** *Data relate to structures solved by X-ray crystallography or electron microscopy.*

It's clearly important to understand the impact of reported resolutions on the reliability of 3D structures. Hence, for each PDB entry, specific details can be found on the *Structure Summary* page, where the full *wwPDB Validation Report* is provided: this includes summaries, with graphical overviews, of geometric issues observed across the constituent molecular entities and their fit with the electron density map, and details (plus magnitudes) of atomic clashes.

Overall, bringing these figures together, PDB's growth statistics reveal that ~132K of the 142K total relate to proteins (**Figure 4, Table 1**), the resolution of ~67% of which is 2.4Å or better; 60% of these entries share >90% sequence identity; and the accumulated total represents just 1,375 unique folds (as defined by CATH).

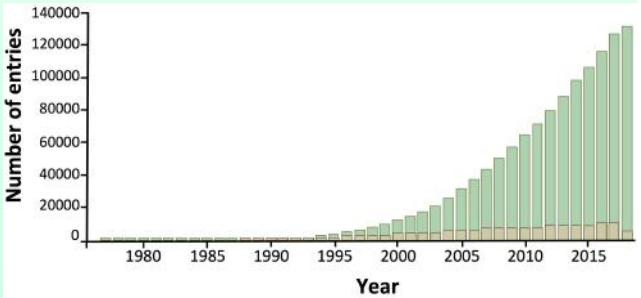

Figure 4 ***PDB protein-only growth statistics, 1 August 2018.*** *The accumulated total number of structures (green), and yearly addition of new structures (gold) are shown.*

KEY TERMS

Diffraction pattern: the pattern of intensities caused by interference of electromagnetic waves (light, X-rays, electron, *etc.*) that have passed through diffracting objects, such as atoms in a crystal

Electron density: a measure of the probability of the presence of an electron at a given location; in diffraction studies, the diffraction pattern gives a probabilistic representation of the locations of electrons

Hydrogen bond: typically, a weak attractive inter- or intramolecular interaction between a hydrogen atom carrying a partial positive charge & a partially negatively charged oxygen or nitrogen atom; they are vital stabilising interactions in protein secondary & tertiary structures, & in nucleic acid secondary structures

Secondary structure: the local, regular organisation of the backbone of a macromolecular structure, stabilised by **hydrogen bonds**: *e.g.*, in proteins, α-helices, β-sheets & β-turns; in nucleic acids, helical stem & loop structures formed by base-pairing interactions

Side-chain: the unique chemical entity attached to the α-carbon of an amino acid molecule

EXERCISES

1 Review the SCOP-defined growth of unique folds: www.rcsb.org/pdb/statistics/contentGrowthChart.do?content=fold-scop. Is the total number of unique folds the same as defined by CATH? If not, why might this be so (think about how SCOP & CATH differ)?

2 Use PDB's generic *Search* box to retrieve the entries for 1f88, 5wkt & 1jfp. Does each entry have a reported resolution? If so, what are they? If not, why might this be so?

3 For each entry, retrieve the full wwPDB Validation Report. What can you infer about the relative quality of these structures?

4 From the *Analyze* menu at the top of the PDB home page, select *Sequence & Structure* alignment. Type 1f88 & 5wkt in the *PDB ID* search boxes, select *jFATCAT-flexible* from the *Comparison Method* pull-down menu & click on the *Align* button. Rotate the image to see how the structures compare. What differences can be seen?

TAKE HOMES

1 The PDB is the primary international archive of the coordinate data of biological macromolecular structures, including those of proteins, nucleic acids & protein-nucleic acid complexes;

2 Originally maintained at the Brookhaven National Laboratory, the PDB has been managed by a consortium of research centres – the Research Collaboratory for Structural Biology (RCSB) – since 1999;

3 The PDB is a member of the wwPDB consortium, which provides an umbrella for coordinating the work of PDBe (Europe), PDBj (Japan) & BMRB (USA), synchronising them with the RCSB PDB;

4 The first release of the PDB was in 1971, with 7 structures; the database now houses coordinate data for >140,000 macromolecules;

5 The PDB holdings are redundant, both in terms of sequence & structural similarity; overall, the database contains ~1,300 unique folds;

6 The PSI was established to hasten the pace of 3D structure determination & to address redundancy issues by targeting novel structures;

7 The number of unique folds entering the PDB did not increase after the launch of the PSI, and has remained at zero since 2012;

8 The PDB may be interrogated by keywords & sequence similarity searches; comprehensive statistics are also available for review;

9 Resolution determines structural quality; overall quality may be examined in the wwPDB Validation Report accompanying each entry.

6 References & further reading

1 **Protein Data Bank.** (1971) *Nature New Biology*, **233**, 223.

2 Kennard, O. *et al.* (1972) **Cambridge Crystallographic Data Centre. I. Bibliographic File.** *J. Chem. Doc.*, **12**(1), 14-19.

3 Kennard, O. (1997) **From private data to public knowledge.** In *The Impact of Electronic Publishing on the Academic Community*. Ian Butterworth, Ed. Published by Portland Press Ltd., London, UK. ISBN 1 85578 122 0

4 Smith, T.F. (1990) **The history of the genetic sequence databases.** *Genomics*, **6**, 701-707.

5 Berman, H.M. *et al.* (2000) **The Protein Data Bank.** *Nucleic Acids Res.*, **28**(1), 235-242.

6 **Protein Data Bank.** (1973) *Acta Crystallogr. sect. B,* **29**, 1746.

7 Bernstein, F.C. *et al.* (1977) **The Protein Data Bank. A computer-based archival file for macromolecular structures**. *J.Mol.Biol.*, **112**(3), 535-542.

8 Berman, H. (2008) **The Protein Data Bank: A historical perspective.** *Foundations of Crystallography*, **64**(1), 88-95.

9 Hobohm, U. *et al.* (1992) **Selection of representative protein data sets.** *Protein Sci.*, **1**(3), 409-417.

10 Berman, H. *et al.* (2003) **Announcing the worldwide Protein Data Bank.** *Nat. Struct. Biol.*, **10**, 980.

11 Young, J.Y. *et al.* (2018) **Worldwide Protein Data Bank biocuration supporting open access to high-quality 3D structural biology data**. *Database*, 10.1093/database/bay002.

12 Burley, S.K. *et al.* (2018) **RCSB PDB: Sustaining a living digital data resource that enables breakthroughs in scientific research & biomedical education.** *Protein Sci.*, **27**, 316-330.

13 Kleywegt, G.J. *et al.* (2018) **Structural biology data archiving - where we are and what lies ahead.** *FEBS Lett.*, **592**, 2153-67.

14 Kinjo, A.R. *et al.* (2017) **Protein Data Bank Japan (PDBj): Updated user interfaces, Resource Description Framework, analysis tools for large structures.** *Nucleic Acids Res.*, **45**(D1): D282-88.

15 Eldon, L. *et al.* (2008) **BioMagResBank.** *Nucleic Acids Res.* **36**, D402-D408.

16 The UniProt Consortium. (2017) **UniProt: the universal protein knowledgebase.** *Nucleic Acids Res.*, **45**(D1), D158-D169.

17 Burley, S.K. *et al.* (1999) **Structural genomics: beyond the human genome project.** *Nat. Genet.*, **23**(2), 151-157.

18 Altschul SF *et al.* (1990) **Basic local alignment search tool.** *J. Mol. Biol.*, **215**(3), 403-410.

19 Dawson NL *et al.* (2017) **CATH: an expanded resource to predict protein function through structure and sequence.** *Nucleic Acids Res.* **45**(D1), D289–D295.

20 Gabanyi, M.J. *et al.* (2011) **The Structural Biology Knowledgebase: a portal to protein structures, sequences, functions, and methods.** *J. Struct. Funct. Genomics*, **12**, 45–54.

21 Haas, J. *et al.* (2013) **The Protein Model Portal - a comprehensive resource for protein structure and model information.** *Database*, 10.1093/database/bat031.

7 Acknowledgements & funding

GOBLET Critical Guides marry ideas from the Higher Apprenticeship specification for college-level students in England (www.contentextra.com/lifesciences/unit12/unit12home.aspx) with the EMBnet Quick Guide concept.

This Guide was developed with the support of a donation from EMBnet to the GOBLET Foundation.

Design concepts and the Guide's front-cover image were contributed by CREACTIVE.

8 Licensing & availability

9 Disclaimer

About the organisations

GOBLET

GOBLET (Global Organisation for Bioinformatics Learning, Education & Training) was established in 2012 to unite, inspire and equip bioinformatics trainers worldwide; its mission, to cultivate the global bioinformatics trainer community, set standards and provide high-quality resources to support learning, education and training.

GOBLET's ethos embraces:

- ***inclusivity**: welcoming all relevant organisations & people*
- ***sharing**: expertise, best practices, materials, resources*
- ***openness**: using Creative Commons Licences*
- ***innovation**: welcoming imaginative ideas & approaches*
- ***tolerance**: transcending national, political, cultural, social & disciplinary boundaries*

Further information about GOBLET and its Training Portal can be found at **www.mygoblet.org** and in the following references:

- Attwood *et al.* (2015) **GOBLET: the Global Organisation for Bioinformatics Learning, Education & Training.** *PLoS Comput. Biol.*, **11**(5), e1004281.
- Corpas *et al.* (2014) **The GOBLET training portal: a global repository of bioinformatics training materials, courses & trainers.** *Bioinformatics*, **31**(1), 140-142.

GOBLET is a not-for-profit foundation, legally registered in the Netherlands: CMBI Radboud University, Nijmegen Medical Centre, Geert Grooteplein 26-28, 6581 GB Nijmegen. For general enquiries, contact **info@mygoblet.org**.

EMBnet

EMBnet, the Global Bioinformatics Network, is a not-for-profit organisation, founded in 1988 as a network of institutions, to establish and maintain bioinformatics services across Europe. As the network grew, its reach expanded beyond European borders, creating an international membership to support and deliver bioinformatics services across the life sciences: **www.embnet.org**.

Since its establishment, a focus of EMBnet's work has been bioinformatics Education and Training (E&T), and the network therefore has a long track record in delivering tutorials and courses worldwide. Perceiving a need to unite and galvanise international E&T activities, EMBnet was one of the principal founders of GOBLET. For more information and general enquiries, contact **info@embnet.org**.

CREACTIVE

CREACTIVE, by Antonio Santovito, specialises in communication and Web marketing, helping its customers to create and manage their online presence: **www.gocreactive.com**.

About the author

Teresa K Attwood (orcid.org/0000-0003-2409-4235)

Teresa (Terri) Attwood is a Professor of Bioinformatics with more than 25 years' experience teaching introductory bioinformatics, in undergraduate and postgraduate degree programmes, and in *ad hoc* courses, workshops and summer schools, in the UK and abroad.

With primary expertise in protein sequence analysis, she created the PRINTS protein family database and co-founded InterPro (her particular interest is in the analysis of G protein-coupled receptors). She has also been involved in the development of software tools for protein sequence analysis, and for improving links between research data and the scientific literature (most notably, Utopia Documents).

She wrote the first introductory bioinformatics text-book; her third book was published in 2016:

- Attwood TK & Parry-Smith DJ. (1999) ***Introduction to Bioinformatics.*** Prentice Hall.
- Higgs P & Attwood TK. (2005) ***Bioinformatics & Molecular Evolution.*** Wiley-Blackwell.
- Attwood TK, Pettifer SR & Thorne D. (2016) ***Bioinformatics challenges at the interface of biology and computer science: Mind the Gap.*** Wiley-Blackwell.

Affiliation

School of Computer Science, The University of Manchester, Oxford Road, Manchester M13 9PL (UK).