Protein Three-dimensional Structure Databases

Vaishali P. Waman[1], Christine Orengo[1], Gerard J. Kleywegt[2], Arthur M. Lesk[3]

1. Institute of Structural and Molecular Biology, University College London, Darwin Building, Gower St, London WC1E 6BT, UK (c.orengo@ucl.ac.uk, v.waman@ucl.ac.uk)
2. European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK. (gerard@ebi.ac.uk)
3. Department of Biochemistry and Molecular Biology and Center for Computational Biology and Bioinformatics, The Pennsylvania State University, University Park PA 16802, USA (aml25@psu.edu)

Keywords: structural biology, data archiving, protein structure, domain analysis, fold classification, protein data bank

Abstract

Databases of three-dimensional structures of proteins provide:

(a) Curated repositories of coordinates of experimentally-determined structures, including extensive metadata, for instance information about provenance, and details about data collection and interpretation, and validation of results.

(b) Information-retrieval tools to allow searching to identify entries of interest, and provide access to them.

(c) Links among databases, especially to databases of amino-acid and genetic sequences, and of protein function; and links to software for analysis of amino-acid sequence and protein structure, and for structure prediction.

(d) Collections of predicted three-dimensional structures of proteins. These will become more and more important after the breakthrough in structure prediction achieved by AlphaFold2.

The single global archive of experimentally determined biomacromolecular structures is the Protein Data Bank (PDB). It is managed by wwPDB, a consortium of five partner institutions: the Protein Data Bank in Europe (PDBe), the Research Collaboratory for Structural Bioinformatics (RCSB), the Protein Data Bank of Japan (PDBj), the BioMagResBank (BMRB) and the Electron Microscopy Data Bank (EMDB). In addition to jointly managing the PDB repository, the individual wwPDB partners offer many tools for analysis of protein and nucleic-acid structures and their complexes, including providing computer-graphic representations. Their collective and individual websites serve as hubs of the community of structural biologists, offering newsletters, support of task groups, training courses, and 'help desks', as well as links to external software.

Many specialised projects are based on the information contained in the PDB. Especially important are SCOP, CATH and ECOD, which present classifications of protein domains.

1 Introduction and Summary

Recently, biology and medicine have become increasingly data-driven. One important component of this enterprise is the storage and distribution of three-dimensional structures (3D). Effective access has required development of databases equipped with tools for information-retrieval and

analysis.

A central focus has been the archive, the Protein Data Bank (PDB) [1, 2]. This is a repository of macromolecular structures, focused on proteins, but also including nucleic acids, and protein and protein-nucleic acid complexes; as well as small-molecule ligands such as NAD, haem, mono- and oligosaccharides, and even stably-bound water molecules. The worldwide PDB – a group of partner organisations – is the custodian, curator, and distributor of these data. The year 2021 marks the fiftieth anniversary of the PDB, recognised by articles and symposia (see https://www.wwpdb.org/pdb50).

The PDB archive sites are rich in links to other databases containing information about proteins. In addition, numerous other databases reformulate and re-present the data. Some are organised around structural and evolutionary relationships; these include SCOP (Structural Classification of Proteins), CATH (Class, Architecture, Topology, Homologous superfamily), and ECOD (Evolutionary Classification Of protein Domains) [3]. Others are 'boutique' databases that focus on specific types of molecules; these include sites presenting G-protein-coupled receptors `https://gpcrdb.org/` and MEROPS, the Database of Proteolytic Enzymes `https://www.ebi.ac.uk/merops/`. Still others focus on specific aspects of the data, such as ligands, or validation. Many sources of software offer tools for access to and analysis of macromolecular structural information. The journal Nucleic Acids Research publishes annual issues on databases, and on webservers, containing but not limited to material about structural data. (`https://academic.oup.com/nar/issue/49/D1`, `https://www.ncbi.nlm.nih.gov/pmc/issues/360280/`.) Protein Science publishes special issues with the theme: 'A compilation of tools for protein science'. (`https://onlinelibrary.wiley.com/toc/1469896x/2020/29/1`)

The archives support research and applications in biology and medicine. The structures also provide training sets for machine-learning algorithms. Box 1 shows typical questions that the PDB might help investigators to pursue.

Box 1. Structural databases support our understanding of the molecular basis of health and disease.

Specific topics include:

- Classification and assignment of protein function; understanding detailed mechanisms of action
- Implications for disease; including effects of mutations
- Interactions among biomacromolecules: the formation and significance of assemblies in normal function and disease
- Identification of proteins with sequences and structures similar to a probe structure
- Species distribution of protein families; evolutionary relationships; phylogenetic trees
- Functional similarity and divergence
- Ligands and modes of binding – support of drug design
- Conformational change in function and regulation

2 Archival databases

The field we now call bioinformatics had its origins in the 1960s and 1970s. In 1965 Margaret O. Dayhoff published the Atlas of Protein Sequence and Structure, collecting the 65(!!) known amino acid sequences of proteins [4]. Dayhoff was also the pioneer in developing computer algorithms for sequence comparison.

Walter C. Hamilton established The Protein Data Bank (PDB) in 1971, originally containing 7 (!!) structures. The first RNA structure was deposited in 1973.

It is difficult for the modern reader to appreciate the infrastructure characterising this era. The original publication of the structure of carboxypeptidase A had the form of a listing, *on paper,* of the atomic coordinates, as Table 3 in a 1970 paper in the Philosophical Transactions of the Royal Society of London [5]. It took a long time – outside specialist communities – before the value of carefully curated and organised data and their availability in computer-readable form became widely accepted. In the early days levels of support for information-oriented projects remained quite small. With no earmarked categories of funding, databases had to compete with – and often were obliged to disguise (not really too strong a word) themselves as – research projects.

With the growth in data productivity, and the recognition of the importance of databases, the field exploded, in the data, and in the computational and human resources devoted to them. Box 2 shows the growth in the contents of the PDB.

The Protein Data Bank at Brookhaven National Laboratories, USA was on its own for 25 years. In 1996, the European Bioinformatics Institute, in Hinxton, Cambridgeshire, UK, initated the Macromolecular Structure Database. Initially a pilot project, it grew and blossomed into the

Protein Data Bank in Europe (PDBe). The Protein Data Bank of Japan (PDBj), at Osaka University, began in 2000. In 1998, the U.S. component was taken over by the Research Collaboratory for Structural Bioinformatics, originally comprising three institutions: Rutgers University, New Jersey; the University of California at San Diego / San Diego Supercomputer Center, in California, and the Center for Advanced Research in Biotechnology / National Institutes for Standards and Technology, at the University of Maryland; recently joined by the University of California at San Francisco (all in the USA).

Box 2. Growth of the Protein Data Bank

Figure 1 shows the time course of the increase in numbers of entries, for the last 30 years. In 1991, the start date of this figure, the PDB contained 694 entries. As of early 2021, the PDB contains 179,000 entries. Growth has been exponential over many years, as presciently predicted by R.E. Dickerson in the late 1970s!

figure_1.eps here

In 2003, the several international institutions agreed to form an umbrella organisation, the worldwide PDB (wwPDB) [6, 7]. In 2006, the Biological Magnetic Resonance Bank (BMRB), now based in Connecticut, USA, also joined. In January 2021, the Electron Microscopy Data Bank (EMDB; also based at EMBL-EBI in the UK) became the latest member. The RCSB, PDBe and PDBj exchange data, co-curate and present the same information in terms of sets of atomic coordinates. However, they evince a very great degree of individuality in terms not only of the 'look-and-feel' of their sites, but the facilities for information retrieval and analysis, the embedded links to other databases, and ancillary features such as newsletters, training courses (see, for example `https://www.ebi.ac.uk/training/online/course/biomacromolecular-structures-introduction-ebi-res`), help desks, and foci on important current topics such as the COVID-19 pandemic.

An important component of interaction of the wwPDB with the community at large is to develop and promulgate standards. In addition to organising Task Forces focussed on specific topics, the community provides input and feedback through the annual wwPDB Advisory Committee meetings.

Of course, databases of macromolecular structures have not grown in a vacuum. They are part of a very large endeavour in biological information storage and organisation. There is very dense network of links between the PDB and the entire 'ecosystem' of biomedical databases.

2.1 Entries

The PDB comprises a collection of entries, each entry corresponding to a single structure determination of a macromolecular structure. Experimental methods are X-ray crystallography – in a few cases, neutron (204 entries) or electron (183 entries) crystallography – NMR spectroscopy, and Electron cryo-microscopy (cryo-EM). (See Box 2.)

Figure 2 analyses the growth of the database, separated according to experimental method.

figure_2.eps here

Figure 3 shows the distribution of the sizes of the entries, measured by numbers of residues. There is a peak about 350 residues, probably covering individual proteins or small oligomers. The entries also include many structures with more than 1900 residues, corresponding to large complexes. For example, the *Rhizobium* arsenite oxidase protein complex contains 3901 residues. Useful for a potential Protein Trivial Pursuit game: The PDB entry with the most residues is the yeast spliceosome, determined by Cryo-EM, containing 38298 residues (entry 3jb9).

figure_3.eps here

Many molecules have been the subject of multiple structure determinations. The PDB contains ∼174,000 protein entries, which correspond to < 55,000 unique UniProt proteins. For example, there are 320 entries for Sperm Whale myoglobin, including mutants. PDBe-KB organises all structures that correspond to a given UniProt entry: For Sperm Whale myoglobin see `https://www.ebi.ac.uk/pdbe/pdbe-kb/proteins/P02185`.

Each independent structure determination corresponds to a separate entry. Different entries can contain the same molecule in different conditions, or in different states of ligation, or showing natural or artificial mutations, or, in the case of crystal structures, at different resolutions. Some molecules appear as the results of separate structure determinations by X-ray crystallography, NMR spectroscopy or electron cryo-microscopy. For different structure determinations of the same molecule under the same conditions – of solvent, temperature and state of ligation – the results are usually very similar. An example: acylphosphatase is a 98-residue enzyme containing two α-helices packed against a $\beta-$sheet. It has been solved both by X-ray crystallography (PDB entry 2acy) and NMR (PDB entry 1aps) (see Figure 4).

figure_4a.jpg figure_4c.jpg figure_4c.jpg here

Many proteins show small or substantial conformational changes as a result of changes in state of ligation. In such cases, different PDB entries containing independent structure determinations of the same molecule may present different conformations. The comparison of oxy- and deoxy-haemoglobin is a classic example of ligand-induced conformational change. Open and closed forms of enzymes provide many other examples. Many of these conformational changes can be described as internal rearrangements of relatively rigid pieces.

In a few cases a protein adopts very different conformations in different contexts. Perhaps the best known example is the prion protein, which can adopt a cellular form (PrP^{C}) that is non-infectious, and a scrapie form (PrP^{Sc}) which causes disease. The two forms have different secondary and tertiary structures. The native and latent states of members of a family of serine-protease inhibitors, the serpins, also show different tertiary structures. The nuclear coactivator binding domain (NCBD) (also called the interferon response binding domain (IbID)) of the creb-binding protein (CBP) shows different structures in complex with the ACTR (acetyltransferase) domain of steroid receptor coactivator p160, and with IRF3 (Interferon Regulatory Factor 3) (see Figure 5). Many proteins can adopt common amyloid structures different from their soluble native conformations.

figure_5a.jpg here figure_5b.jpg here

Even a single entry may contain the coordinates of multiple copies of the same protein with different structures. For X-ray structures, this is in some cases the result of different crystal-packing environments. But not always: An mRNA capping enzyme adopts open and closed conformations – both of which appear, in the same state of ligation, in the same asymmetric unit of a crystal [1ckm]. The iron-sulphur protein IscU [2z7e] is another example. A particularly interesting case is the packing of coat proteins in icosahedral viruses, such as Tomato Bushy Stunt Virus, in which multiple copies of a single protein adopt three different conformations to satisfy the requirements of the symmetry of the particle. For NMR structures, entries generally contain multiple models of a structure, all more or less equally consistent with the experimental data (see Figure 4b).

Many proteins contain intrinsically-disordered regions, or even are entirely disordered [9], (`https://www.nature.com/subjects/intrinsically-disordered-proteins`). Needless to say, coordinates of disordered atoms do not appear in the PDB entries.

2.1.1 The PDB files

On the PDBe, RCSB, and PDBj web sites, each entry has a separate page, with its own URL. Each site contains a link allowing download of the entry. These files, the result of curation and validation, are *common* to PDBe, RCSB, and PDBj. However, the entry pages offered by the different wwPDB partner sites differ in appearance, facilities, and links. Of course there is substantial overlap; for instance, each offers visualization of the structure.

The *raison d'être* of the entry is the set of coordinates. These may specify the atomic positions in:

- A protein – for instance, hen egg white lysozyme
- A nucleic acid – for instance, transfer RNA
- A complex of several proteins – increasingly including large and complex molecular assemblies, as a result of recent advances in electron cryo-microscopy
- A complex of protein and nucleic acid, including relatively simple ones such as the antennapedia homeodomain-DNA complex, containing a 62-residue protein and a 15-basepair fragment of DNA; and large ribonucleoprotein assemblies, *e.g.,* the ribosome.
- In addition to macromolecules, the entries may contain the coordinates of ligands of many chemical types, such as atomic ions (Cl^-, Mg^{2+}), SO_4^{2-}, common cofactors such as NAD and FAD, the haem groups, *e.g.,* in globins and cytochromes *c,* mono- and oligosaccharides, and also including stably-bound waters.

Entries also contain information about the source of the molecule; experimental details about the structure determination, including lists of missing residues if any; in the cases of proteins, secondary structure (helix and sheet) assignments, and links to publications.

Once posted, it is rare for an entry to be altered, other than to correct obvious and trivial typographical errors. (In one entry, for a cytochrome c, the molecule name in the TITLE record was for a long time misspelt as cytrochrome; although this would not confuse a human reader, it caused problems for information-retrieval software based on free-text searches.) Updated versions of entries, by the original authors, are now possible, following further rebuilding and refinement *based on the same experimental data.* In other cases, sometimes a redetermination of a structure, for instance at higher resolution, leads to replacement of an entry by the new result that supersedes it. The removed entries are archived and accessible if necessary. In some rather strange cases, structures were deposited that were incorrect because the crystallographer had forced a model into an electron-density map of the wrong enantiomorph; these were subsequently corrected. In a few cases, entries that are entirely incorrect have been removed when the journal articles reporting them were withdrawn [10].

2.2 Deposition of a PDB entry

Experimental structure determinations of macromolecular data enter the PDB via a common system shared by the partner institutions, OneDep (see Figure 6) [11]. The country of origin of the submission determines which institution will be responsible for its processing (see Figure 3 in reference [11].)

figure_6.eps here

An entry submitted to the Protein Data Bank is subjected to an 'entrance exam', checking for errors, completeness and consistency, and editing if necessary to convert the material into standard format and nomenclature.

It is now mandatory to submit supporting experimental data, in addition to a coordinate set (Table 1). For instance, for a structure determination by X-ray crystallography it is necessary to deposit structure-factor data. This not only supports the validation process, but allows improvement of the structure after advances in the interpretation of the data (see PDBREDO, below).

The question of access to scientific data is a very old one. Late in the seventeenth century, Isaac Newton demanded access to data collected by the Astronomer Royal, John Flamsteed, in order to prepare a new edition of *Principia*. In 1695, Newton wrote to Flamsteed: '... these and almost all your communications will be useless to me unless you can propose some practicable way or other of supplying me with Observations ... *I want not your calculations but your Observations* only.' (italics by current authors) Flamsteed refused, claiming ownership of the data despite their having been collected while he occupied an official government post.

For contemporary macromolecular structures, it was also not trivial to impose requirements for deposition of data [12, 13]. (There were many Flamsteeds in the protein-crystallography community.) The International Union of Crystallography (IUCr), through its Commission on Biological Macromolecules, published guidelines [14]. Originally, these were solely advisory. Genuine pressure on scientists could be applied *via* requirements for publications in journals, and for grant support. Now, PDB deposition requires supporting experimental data (see Table 1), and many journals require official PDB validation reports to support review of a submitted article.

Table 1 here

In addition to deposition of derived experimental data in the PDB, other archives include The Integrated Resource for Reproducibility in Macromolecular Crystallography (IRRMC) (`www.proteindiffraction.org`), SBGRID (`https://data.sbgrid.org/`), and (for Australian datasets) TARDIS Raw Diffraction Data Archive (`researchdata.edu.au/tardis/14901`). For electron cryo-microscopy, the Electron Microscopy Public Image Archive (EMPIAR) (`www.ebi.ac.uk/pdbe/emdb/empiar/`) collects and distributes the raw images from which three-dimensional structures are derived.

For X-ray crystal structures, the Uppsala Electron-Density Server presented electron-density maps. The PDBe has now taken over this task, providing maps for crystallographic and EM structures: `https://www.ebi.ac.uk/pdbe/about/news/pdbe-brings-electron-density-viewing-masses`. If, for example, a series of high atomic displacement parameters (ADPs) (often called B-factors) raises suspicions (ADPs are parameters suggesting the precision of the coordinates of individual atoms), inspection of the maps can allay or reinforce them.

2.3 Validation

The PDB subjects depositions of new entries to several kinds of checks [15]. These are based on recommendations from community input – Validation Task Forces (`https://www.wwpdb.org/task/validation-task-forces`). Some of the same criteria, such as stereochemical checks, apply all structures based on different experimental methods; others are experimental-method dependent. A software pipeline implementing the validation procedures produces reports, accessible from the entry pages in the wwPDB sites. The web site `https://validate.wwpdb.org` allows users to upload their own structures to this software – for instance, a potential depositor might wish to pre-check a structure in preparation for submission.

Checks include:

1. Stereochemical tests independent of the associated experimental data; for instance:
   - Bond length and bond angle outliers from expected values
   - Outliers in the Ramachandran plot
   - Steric clashes between non-bonded atoms.
   - Outliers of sidechains deviating from standard rotamer conformations
   - It has been pointed out by crystallographers – *very* emphatically – that outliers do not necessarily signal errors in structure determinations. (Conversely, non-outliers also may or may not be errors.)

     (Outliers are *usually* either errors, or in some cases may be unusual but interesting features meriting closer scrutiny. However, such features are credible only if the experimental data *clearly* support them.)

2. Analysis of the experimental data independent of the coordinate set. This is method dependent. For example, for X-ray crystal structures, the Wilson B factor (derived from the variation of average reflection intensity with $\sin 2\theta/\lambda$) indicates the degree of order in the crystal.

3. Analysis of the fit between the atomic coordinates and the experimental data. This is also method dependent; for X-ray crystal structures, values include the R-factor and $\mathrm{R_{free}}$. The R-factor is a measure of the agreement between observed (obs) structure factor (F) amplitudes and the corresponding quantities calculated from the model (calc): $R = \frac{\sum ||F_{\mathrm{obs}}|-|F_{\mathrm{calc}}||}{\sum |F_{\mathrm{obs}}|}$. $\mathrm{R_{free}}$ is the same sum, but over only a test set consisting of $\sim 1-10\%$ of the data which are *excluded* from refinement of the coordinate set. Also for crystal structures, the RSR Z-score measures the quality of the fit between each residue and the electron-density map (computed from phases derived from the model).

Suppose the PDB contains two or more structures of the same molecule, and you wish to choose the best one for some application, such as studying evolution in a family of proteins, interpreting an enzyme mechanism, or docking potential ligands for drug design. Is there a way to extract from the validation statistics a single number to govern your choice? The simple answer is no. However, the PDBe does compute a composite 'score' that takes into account model validation, map-model validation, and resolution; the entries returned by a search can be sorted according to this quality score.

For X-ray crystal structures, the resolution *in principle* imposes limits on the precision of the derived coordinates. However, it is not always true that the higher the resolution, the higher the quality of the structure.

For NMR structures, there is no commonly-accepted and widely-used equivalent of resolution. The program Resolution by Proxy (ResProx) computes an equivalent of resolution, based solely on the coordinate data [16]. (For crystal structures, ResProx values correlate very well with reported resolution values.)

Measures of quality of electron cryo-microscopy structures are the subjects of current debate [17, 18]). (A report from the EM Validation Task Force is expected later this year.)

2.3.1 PDBREPORT

Other projects offering 'health checks' of deposited proteins include PDBREPORT (`https://swift.cmbi.umcn.nl/gv/pdbreport/`) [19].

The PDBREPORT database contains the results of validation software, WHAT_CHECK, applied to each entry in the Protein Data Bank. The kinds of problems flagged by PDBREPORT include simple 'howlers', and more subtle anomalies. The program tests the validity and consistency of the format, and also analyses the structures, detecting outliers in stereochemical properties, such as bond lengths or angles, and checks for consistency in hydrogen-bonding patterns.

2.3.2 PDBREDO

Progress in X-ray structure determination has depended in part on advances in data collection, especially the use of synchrotrons, but also in the software that converts the measured data into a structure. Applying new software can improve the results. The PDBREDO project produces re-refined models for most X-ray entries in the PDB; usually with significantly improved validation statistics. The new results are also unbiased by choices made by individual scientists during the original structure determinations [20, 21].

Figure 7 shows typical changes that PDBREDO produces. The PDB-REDO databank `https://pdb-redo.eu` contains 170000 entries.

figure_7.jpg here

3 Structure of the data, and the surrounding information ecosystem

The goal is adherence to the FAIR principles of data archive infrastructure: data must be Findable, Accessible, Interoperable and Reusable [23].

The organisation of the PDB is based on entries – one file per structure determination. (Note that the PDB itself is an archive of files, NOT a database, but it provides a basis for database design and implementation.) The wwPDB websites present an 'vestibule' page for each entry, containing a summary of essential information, images, *etc.* These pages provide access to the entry files, and copious associated information. The pages differ in appearance and links among the wwPDB partner sites. Figure 8 illustrates an entry for Sperm Whale myoglobin, 1mbo, in PDBe [24]. Many items on this page are links, to detailed annotations – and search engines supporting identification of other entries with shared selected annotations – and to other databases, notably

UniProtKB. UniprotKB is an umbrella database about proteins, organised around amino-acid sequences (that is, not limited to proteins for which three-dimensional structures have been determined) comprehensively collecting available annotations, including but not limited to function, and providing search engines for family relationships. Other links allow entry downloads from within the browser. (See also `https://www.wwpdb.org/ftp/pdb-ftp-sites`).

Close-coupled to the PDBe component of the wwPDB, the PDBe-KB (Protein Data Bank in Europe – Knowledge Base) collates structural and functional annotations of macromolecular structures, both literature-derived and computationally predicted; a major goal is to set macromolecular structure data in their biological context. One can think of links from the PDB to UniprotKB as portals to more comprehensive annotations of PDB entries, and links from UniprotKB to PDBe-KB as portals to structural information available for UniprotKB entries. Because PDBe-KB is organised according to UniProt ID, one PDBe-KB entry may refer to many PDB entries, or even to different parts of a molecule.

A number of derived databases present hierarchical structural classifications of domains within PDB entries – notably SCOP, CATH and ECOD (see section 4). Several other projects seek to embed the Protein Data Bank in more general databases that link structure with other properties of proteins. These include:

- canSAR [25] collects multidisciplinary data on relevant proteins including structure, ligands and other interactions, and clinical annotations, in support of drug discovery.
- The Elixir 3D-Bioinfo organisation [26] has organisational goals to strengthen interactions in the European research communities for structural biology and related fields, and stimulate infrastructure development.

figure_8.jpg here

3.1 Selection and retrieval of entries

The wwPDB partner sites contain powerful search engines allowing users to *identify* entries of interest [24]. There are two basic approaches:

- **Query by property:** In addition to the name of the molecule and keywords in its description, users can impose criteria based on a large number and variety of features, including: Scientific Name of Source Organism; Taxonomic Class of Source Organism (*e.g.,* Archaea); Polymer Entity Type (Protein, Nucleic acid); Release Date; Experimental Method, for X-ray structures: Resolution Range; UniProt identifiers, Enzyme Commission Classification (highest level class, *e.g.,* Hydrolases); Gene Ontology Classification terms, names of ligands, authors of related articles and other publication details – and many others.
- **Query by similarity in sequence or structure** to a probe object. There are various tools for searching the PDB for structures with sequence similarity to a submitted *sequence*. These include general sequence-search tools: BLAST; PSI-BLAST; and Hidden Markov Models – for instance the program HMMER [29–33]. In each case one would select a list of amino-acid sequences of PDB entries as the search space. Typically pairwise sequence similarity is measured in terms of the % identical residues in an optimal sequence alignment, or scores based on a substitution matrix such as BLOSUM-62, or E-value. (The Expect-value, or E-value, of an item returned by a search estimates the number of hits one can 'expect' to

find by chance in the database search. The E-value depends on the size of the database searched.) The RCSB offers the MMseqs2 method for similarity searching [34] (instead of BLAST), achieving ∼11 times faster performance.

Most entries in PDBe are linked to a corresponding entry in PDBe-KB (the sequence must be in UniProt) which contains a pre-stored list of other PDB entries with similar sequences. In RCSB and PDBj, entries are linked to search engines for other entries similar in sequence, motifs, structure, and common ligands.

Searches for similar *structures* are also possible. Often, pairwise structure similarity is measured in terms of the root-mean-square distance of Cα atoms of corresponding residues derived from a structure-based alignment, after optimal superposition. Given two sets of corresponding atoms, with coordinates $x_i, y_i, z_i, i = 1, \ldots n$, and $X_i, Y_i, Z_i, i = 1, \ldots n$, to measure structural similarity compute the root-mean-square distance (r.m.s.d.): $\sqrt{\sum_{i=1}^{n}[(x_i - X_i)^2 + (y_i - Y_i)^2 + (z_i - Z_i)^2]/n}$ after optimal superposition.

PDBeFold contains a set of relevant tools, built around the structural alignment program SSM [35]. RCSB has a Structure Similarity option in its Query Builder, and also sets of Systematic Pre-calculated Protein Structure Alignments. For structure similarity searching, the RCSB has developed a computationally efficient method based on Zernike polynomials [36]. (`http://shape.rcsb.org/`) The RCSB also offers sequence motif searching using simple queries, or PROSITE patterns or regular expressions. PDBj offers a Sequence navigator page `https://pdbj.org/seqnavi`, which allows selecting a chain from a PDB entry and finding similar structures within the PDB; users may view a structural superposition interactively in three dimensions. A related feature, SeSAW, identifies motifs similar in sequence and structure to a query protein. `https://sysimm.org/sesaw.2.0/` The PDBj also offers a shape-similarity search tool, Omokage.

Many other programs, independent of the PDB, also carry out structure-similarity searches and structure-based sequence alignments.

The wwPDB partners present several high-performance application programming interfaces for information retrieval. The RCSB recently introduced the 'Data API' and the 'Search API'. (`https://www.rcsb.org/pages/webservices`) The Data API comprises interfaces: REST-API and GraphQL-APIs. GraphQL offers more flexible data access to any level of hierarchy in the full PDB data schema. The Search API allows very powerful and general queries. The PDBe and PDBj also have APIs and graph databases.

3.2 Software for analysis

With the growing maturity and availability of software that supports research in bioinformatics in general and structural bioinformatics in particular, the 'ramp' from data retrieval to analysis is becoming smoother. Today it is rare to follow the old regime of downloading datasets to a local computer, and 'rolling one's own' programs to process them. The databanks themselves offer online facilities for many operations, and links to many other programs. These differ among PDBe, RCSB, and PDBj.

For example PDBe provides:

- **PDBeFold:** pairwise and multiple three-dimensional alignment and superposition. (`https://www.ebi.ac.uk/msd-srv/ssm/`)
- **PDBePISA:** Explore macromolecular interfaces; predict quarternary structure. (`https:`

//www.ebi.ac.uk/msd-srv/prot_int/pistart.html)

- **wwPDB validation server:** Generate X-ray structure validation reports to assess the quality of (hopefully!) successively-improved models and to identify potential problems that need to be addressed; intended to be useful to structure-determination projects, in preparation for submission of results to PDB. (https://validate-pdbe.wwpdb.org/) All wwPDB partners provide access to the validation server, which also treats for NMR and EM structures.

PDBj offers:

- **Yorodumi:** Interactive display of 3D structures of biological molecules (H. Suzuki https://pdbj.org/emnavi/quick.php) Yorodumi is Japanese for 'Ten thousand views'
- **ASH (Alignment of Structural Homologs):** structure alignment (D.M. Standley, H. Toh & H. Nakamura https://sysimm.org/ash_service/)
- **MAFFTash:** multiple sequence alignments from sequences and structures (K. Katoh & H. Toh http://sysimm.org/MAFFTash//)
- **gmfit:** superposition of atomic models and 3D density maps (T. Kawabata https://pdbj.org/gmfit/)
- **CRNPRED:** prediction of secondary structure and contact orders (A.R. Kinjo & K. Nishikawa https://pdbj.org/crnpred/)
- **Spanner:** thread an amino-acid sequence into a PDB structure (M. Lis, T. Kim, J. Sarmiento, D. Kuroda, H. Dinh, A.R. Kinjo, S. Devadas, H. Nakamura & D.M. Stanley https://pdbj.org/spanner)
- **SFAS:** from submitted amino-acid sequence, predict secondary structure and disordered regions, or search for homologues via HHPred (comparison of profile hidden Markov models) (https://pdbj.org/sfas/)
- **HOMCOS:** Finding and modeling of 3D structures of complexes (N. Fukuhara & T. Kawabata http://homcos.pdbj.org/)

The wwPDB web sites offer interactive visualisation software. For RCSB and PDBe, the current default viewer is Mol*. PDBj uses molmil.

Many other institutions offer suites of useful programs; see for instance the web server of the Monash University Laboratory of Computational Biology, under the direction of Dr Arun Konagurthu. This includes:

- **seqMMLigner:** pairwise sequence alignment [37] (https://lcb.infotech.monash.edu/seqmmligner/)
- **MMLigner:** pairwise structure alignment [38] (https://lcb.infotech.monash.edu/mmligner/)
- **MUSTANG:** multiple structure alignment [39] (https://lcb.infotech.monash.edu/mustang)

- **Super:** rapid screening of the entire (up-to-date) PDB to identify oligopeptide fragments with backbone structure similar to a probe fragment (or to two fragments with a prespecified gap between them) [40] (`https://lcb.infotech.monash.edu.au/super/`)
- **SST:** assignment of secondary structure to protein coordinate set [41] (`https://lcb.infotech.monash.edu/sstweb2/Submission_page.html`)
- **Proçodic:** dissection of a structure into components of a fixed library of substructures, an architectural 'basis set' of observered protein structures [42] (`https://lcb.infotech.monash.edu/prosodic`)

3.3 Formats of individual entries

Although the core of a PDB file is the coordinate set, substantial additional information describing the structure determination and its subject also appears. This includes information about the source and function of the molecule(s) that are the subject of the structure determination, the scientists responsible, reference to a journal publication (in almost all cases), and technical information about the project; for instance, for X-ray structure determinations, the resolution range and refinement statistics. Some interpretative annotations include secondary-structure assignments, for entries containing proteins.

The 'classical' PDB format was developed almost fifty years ago, and was limited by contemporary technology. In particular, files were stored on 'punched cards', 80 columns in width. (For the new generation of scientists who have never encountered a punched card: see `https://twobithistory.org/2018/06/23/ibm-029-card-punch.html`) Data were stored line-by-line – one card to a line – and internal references but not electronic links were possible.

Each line began with a six (or fewer)-letter keyword specifying its contents. In particular, coordinate data lines began with ATOM (for protein atoms) or HETATM (for ligands, including water.) To give an impression of the format, Figure 9 shows a few lines extracted from the file for a structure of Sperm Whale Myoglobin, PDB entry 1mbo. (The complete file is 2040 lines long. This is JUST slightly more than one standard box of 2000 punch cards.)

figure_9.eps here

It became clear that this format could not support modern database infrastructure. Moreover, it cannot capture the very large and complex structures that can be determined now. It survives because of the immense amount of legacy software based on it.

An International Union of Crystallography (IUCr) Working Party on Crystallographic Information developed a new format, called CIF (for Crystallographic Information File). Developed originally for small-molecule data, it was adopted in 1990, and expanded for macromolecules as mmCIF [43]. This is now the standard. See also: `https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/beginner%E2%80%99s-guide-to-pdb-structures-and-the-pdbx-mmcif-format`

For comparison, here is an extract from the mmCIF file for the PDB entry, 1mbo. An identifying phrase (not limited to a 6-letter keyword), beginning with an underscore, introduces each set of data items. Excluding lines containing coordinates of atoms, the PDB file for 1mbo contains 429 lines. The mmCIF file for the same entry contains, again exclusive of ATOM coordinate lines, 1622 lines. Of course there is strong motivation to pack information into each line if

the file is to be stored on punched cards.* The relative character counts (after truncating trailing spaces) are: PDB file for 1mbo: 145677; mmCIF file: 198108, a smaller discrepancy.

*A (very rough) back-of-the-envelope calculation suggests that to store ∼50 PDB entries on punched cards in mmCIF format rather than old PDB format would cost one 8-inch-diameter tree.

```
data_1MBO
#
_entry.id   1MBO
#
_audit_conform.dict_name       mmcif_pdbx.dic
_audit_conform.dict_version    5.279
_audit_conform.dict_location   http://mmcif.pdb.org/dictionaries/ascii/mmcif_pdbx.dic
#
loop_
_database_2.database_id
_database_2.database_code
PDB   1MBO
WWPDB D_1000174929
#
_pdbx_database_status.status_code                     REL
_pdbx_database_status.entry_id                        1MBO
_pdbx_database_status.recvd_initial_deposition_date   1981-08-27
_pdbx_database_status.deposit_site                    ?
_pdbx_database_status.process_site                    ?
_pdbx_database_status.status_code_sf                  REL
_pdbx_database_status.status_code_mr                  ?
_pdbx_database_status.SG_entry                        ?
_pdbx_database_status.status_code_cs                  ?
_pdbx_database_status.methods_development_category    ?
_pdbx_database_status.pdb_format_compatible           Y
#
_audit_author.name           'Phillips, S.E.V.'
_audit_author.pdbx_ordinal   1
#
...
_entity.details
1 polymer     man MYOGLOBIN                         17234.951 1   ? ? ? ?
2 non-polymer syn 'SULFATE ION'                     96.063    1   ? ? ? ?
3 non-polymer syn 'PROTOPORPHYRIN IX CONTAINING FE' 616.487   1   ? ? ? ?
4 non-polymer syn 'OXYGEN MOLECULE'                 31.999    1   ? ? ? ?
5 water       nat water                             18.015    334 ? ? ? ?
#
...
_entity_poly.pdbx_seq_one_letter_code
;VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASEDLKKHGVTVLTALGAILKKKG
HHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYKELGYQG
;
...
_struct_conf.pdbx_PDB_helix_length
HELX_P HELX_P1 A SER A 3   ? GLU A 18  ? SER A 3   GLU A 18  1 ? 16
HELX_P HELX_P2 B ASP A 20  ? SER A 35  ? ASP A 20  SER A 35  1 ? 16
HELX_P HELX_P3 C HIS A 36  ? LYS A 42  ? HIS A 36  LYS A 42  1 ? 7
HELX_P HELX_P4 D THR A 51  ? ALA A 57  ? THR A 51  ALA A 57  1 ? 7
HELX_P HELX_P5 E SER A 58  ? LYS A 77  ? SER A 58  LYS A 77  1 ? 20
HELX_P HELX_P6 F LEU A 86  ? THR A 95  ? LEU A 86  THR A 95  1 ? 10
HELX_P HELX_P7 G PRO A 100 ? ARG A 118 ? PRO A 100 ARG A 118 1 ? 19
HELX_P HELX_P8 H GLY A 124 ? LEU A 149 ? GLY A 124 LEU A 149 1 ? 26
#
...
_atom_sites_footnote.text
1
;INSPECTION OF MAPS AT VARIOUS STAGES OF REFINEMENT INDICATED THAT SEVERAL SIDE-CHAINS AND THE C- AND N-TERMINAL RESIDUES WERE DISORDERED TO
;
2 'HOH 305 IS CLOSE TO ATOM O1 OF HEM 546. SEE PAPER CITED AS JRNL REFERENCE ABOVE.'
#
...
_atom_site.pdbx_PDB_model_num
ATOM   1    N  N   . VAL A 1 1   ? -0.686 14.852 16.090  1.00 42.31 ? 1   VAL A N   1
ATOM   2    C  CA  . VAL A 1 1   ? 0.621  15.543 16.081  1.00 53.86 ? 1   VAL A CA  1
ATOM   3    C  C   . VAL A 1 1   ? 1.823  14.549 16.200  1.00 27.97 ? 1   VAL A C   1
ATOM   4    O  O   . VAL A 1 1   ? 1.628  13.369 15.953  1.00 23.24 ? 1   VAL A O   1
ATOM   5    C  CB  . VAL A 1 1   ? 0.521  16.582 17.208  1.00 36.55 ? 1   VAL A CB  1
ATOM   6    C  CG1 . VAL A 1 1   ? -0.444 17.695 16.811  1.00 35.23 ? 1   VAL A CG1 1
ATOM   7    C  CG2 . VAL A 1 1   ? 0.119  15.972 18.559  1.00 53.38 ? 1   VAL A CG2 1
ATOM   8    N  N   . LEU A 1 2   ? 3.112  14.944 16.272  1.00 13.35 ? 2   LEU A N   1
ATOM   9    C  CA  . LEU A 1 2   ? 4.140  13.989 16.663  1.00 16.70 ? 2   LEU A CA  1
ATOM   10   C  C   . LEU A 1 2   ? 3.978  13.813 18.182  1.00 18.20 ? 2   LEU A C   1
ATOM   11   O  O   . LEU A 1 2   ? 3.770  14.833 18.864  1.00 17.87 ? 2   LEU A O   1
ATOM   12   C  CB  . LEU A 1 2   ? 5.553  14.489 16.342  1.00 8.27  ? 2   LEU A CB  1
ATOM   13   C  CG  . LEU A 1 2   ? 6.119  14.183 14.949  1.00 12.85 ? 2   LEU A CG  1
ATOM   14   C  CD1 . LEU A 1 2   ? 5.542  15.081 13.861  1.00 15.27 ? 2   LEU A CD1 1
ATOM   15   C  CD2 . LEU A 1 2   ? 7.651  14.312 14.995  1.00 20.18 ? 2   LEU A CD2 1
[additional atoms for residues 3-153]
...
HETATM 1233 C  CHA . HEM C 3 .   ? 16.907 31.140 0.448   1.00 4.00  ? 155 HEM A CHA 1
HETATM 1234 C  CHB . HEM C 3 .   ? 15.562 27.978 3.977   1.00 4.00  ? 155 HEM A CHB 1
HETATM 1235 C  CHC . HEM C 3 .   ? 14.124 24.760 0.639   1.00 7.11  ? 155 HEM A CHC 1
HETATM 1236 C  CHD . HEM C 3 .   ? 15.416 28.028 -2.786  1.00 7.65  ? 155 HEM A CHD 1
HETATM 1237 C  C1A . HEM C 3 .   ? 16.738 30.484 1.707   1.00 5.70  ? 155 HEM A C1A 1
HETATM 1238 C  C2A . HEM C 3 .   ? 16.913 31.113 2.978   1.00 7.72  ? 155 HEM A C2A 1
[additional atoms for haem group]
```

Programs are available to interconvert PDB and mmCIF formats. The site `https://mmcif.wwpdb.org` provides a variety of resources.

4 Structure classification databases

Ernest Rutherford famously said: 'All science is either physics or stamp collecting.' One of us has retorted: '... the study of protein structure combines the best elements of both' [8]. Such a large and varied corpus of data as protein structures presents the challenge of how to classify them. Consensus has emerged that the proper objects to compare are protein domains [44].

What is a domain? There is agreement that a domain is a compact substructure within a protein that appears to have independent stability. Some scientists require also that for a substructure to be considered a domain it must appear in different protein structures associated with different partner domains. Of course for many proteins the entire structure forms one compact entity and these are regarded as single-domain proteins.

Domain-based structural classification projects (see Table 2), focus on understanding evolutionary relationships in proteins and associated functional divergence. To allow a comprehensive classification of protein-folding patterns, the highest levels group proteins according to structural similarity, independent of evolutionary relationship; lower levels classify based on evolutionary divergence.

Table 2 here

SCOP (**S**tructural **C**lassification **O**f **P**roteins) and CATH (**C**lass **A**rchitecture **T**opology **H**omologous Superfamily) were the first initiatives for comprehensive structure-based classification of protein domains [45–47]. When they began, there were ∼3000 structures in the PDB; now, in early 2021 there are more than 179,000 – growth has been exponential (see Figure 1). Over the past 25 years, classification schemes and annotations from SCOP and CATH have supported many investigations of protein structure, function, and evolution; and applications to prediction of protein structures, functional sites and protein-protein interactions (reviewed in [3, 48, 49].)

A second generation of entries to the field, 20 years later, in 2014, include: Brenner and co-workers updated SCOP, and established the SCOPe resource [50]. The SCOPe website provides access to all releases of SCOP and Astral databases that feature stable identifiers. Murzin's group has developed a new database, SCOP2, with many major changes. Grishin and coworkers produced the ECOD (**E**volutionary **C**lassification **O**f Protein **D**omains).

With the development of SCOP2, the original SCOP database is no longer updated and maintained. The latest version, SCOP1.75, is a legacy database. Its coverage of the PDB is incomplete, in terms of PDB entries included; however, it can be estimated that SCOP1.75 contains examples of approximately 80% of known domain 'fold space' (admittedly not a statement that allows for precise definition).

Here we will describe the groupings in CATH, SCOP, SCOP2 and ECOD, and the strategies used for classification. Perhaps the most important development is the move, in SCOP2, away from strictly hierarchical classifications. This arose from the need to accommodate the considerable

structural divergence observed in some of the most highly populated superfamilies, and the greater variety of important relationships among proteins. For example, in the SCOP HUP superfamily (HIGH-signature proteins, UspA superfamily and PP-ATPases), some relatives can vary in size by up to 5-fold, at which point it is difficult to assign them to the same fold group. Indeed, SCOP places some of these extremely divergent homologues in different superfamilies. To solve this, SCOP2 identifies structural relationships between superfamilies and highlights structural motifs shared between them. As these phenomena became more apparent, CATH refined the definition of fold group, or T-level, to reflect similarity in the structural cores of the domains, which is typically highly conserved even in very divergent homologues [51]. CATH also identifies Structurally Similar Groups (SSGs) within superfamilies. In ECOD, the X-level groups superfamilies in which domains have structural similarity in the core, whereas in ECOD the T-level within the superfamily sub-classifies relatives having more global levels of structural similarity.

Table 3 reports the number of entries in each level of each of the structural classification databases (as of February, 2021).

Table 3 here

To classify protein domains on the basis of structure and function, two problems must be solved: (1) identification of domains within proteins and (2) detection of homology among domains:

(1) Databases use a combination of automatic and manual procedures for dissection of a protein structure into domains.

In CATH, an automated process, AutoChop, scans query structures against chains that have already been dissected into their constituent domains in CATH. AutoChop includes ChopClose [52] which uses the SSAP (**S**econdary **S**tructure **A**lignment **P**rogram) algorithm [53, 54]. If a match is found, the alignment induces the dissection of the query chain. For chains not treated successfully by AutoChop, a manual curation procedure identifies domains, making use of several software tools, plus comments appearing in the scientific literature.

SCOP and SCOP2 also use software tools for obvious cases, but, more than other projects, depend on personal judgement, that of Alexei Murzin. Murzin has extremely unusual perceptual gifts – coupled with profound understanding of protein architecture. In consequence, SCOP and SCOP2, to a much greater extent than other databases, rely on his curatorial expertise. (Indeed, for subtle and tricky problems, Murzin is regarded in the field as 'the court of last resort'.)

The SCOPe domain-classification protocol first uses a sequence-based similarity scan against a database of SEQRES-based sequence domains from SCOP and SCOPe [49, 50]. Thus, new query protein chains are scanned against the SCOPe database using BLAST and significant matches identified based on an E-value cutoff of 10^{-4} and maximum alignment coverage. If successful, the domains in the query protein chain are assigned based on the top-ranking BLAST-based alignments. If automated classification is unsuccessful, there is recourse to manual curation.

ECOD has adopted structures from SCOP v1.75. For structures not contained in SCOP v1.75, ECOD applies a combination of sequence and structural homology detection software, backed up by visual inspection and comparison, plus information from the scientific literature [55].

(2) Homology detection is quite a tricky problem. Among very close relatives it may be considered as

obvious; conversely, for pairs of proteins with no visible signal at all of relationship, lack of homology must be assumed. It is well-known that structural similarity survives evolutionary divergence far more robustly than sequence similarity. But even many cases of tantalising structure similarity are difficult to resolve.

CATH uses automated approaches, and manual curation for validation of distant homologues. Homologues must meet two out of the following three criteria:

1. *structure similarity,* determined by SSAP
2. *sequence similarity,* based on sequence identity or E-value from Hidden Markov Model (HMM)-based methods
3. *functional similarity,* based on evidence in the literature

The criterion for homology in SCOP is: (1) proteins with >30% residue identity in aligned sequence, or (2) proteins with less-similar sequences, but for which the structures and function are very similar.

ECOD builds on superfamily data from SCOP but also uses more automated approaches to detect remote homologues [56].

4.1 The CATH database

CATH is an acronym for the four levels in the hierarchical classification scheme (**C**lass, **A**rchitecture, **T**opology, **H**omologous Superfamily) [57]. The CATH database and website were established in 1994 [58]; since then, CATH has been maintained regularly and kept up to date [59–61].

To focus on the best-determined structures, CATH retains only PDB entries that are: experimentally-determined protein structures, with (i) length > 40 amino acid residues, (ii) resolution (of crystal structures) $\leq$ 4 Å, and (iii) >70% of side chains resolved.

4.1.1 Hierarchical groupings in CATH

CATH classifies protein domains into four major hierarchical categories: (1) Class (C-level), (2) Architecture (A-level), (3) Topology (Fold or T-level), (4) Homologous Superfamily (H-level) [58]. CATH subclusters Homologous Superfamilies into distinct Functional Families (FunFams), based on predicted similarity of function [62].

- **Classes / C-level in CATH**: Class is the most general level in CATH. Levitt & Chothia, in 1976, first proposed classifying proteins into structural classes, to reflect secondary structure composition [63]. They grouped proteins into four major classes: all-α, all-β, $\alpha\beta$ (proteins containing both α-helices and β-sheets, with helical and sheet regions segregated to different parts of the structure; β-sheets usually antiparallel) and $\alpha + \beta$ (containing β-α-β supersecondary structure units, as in the NAD-binding fold). CATH classifies protein domains into mainly-α (Class 1), mainly-β (Class 2), α-β (Class 3, which combines α/β and $\alpha+\beta$), domains with few or even no secondary structures (Class 4), and multidomain proteins (Class 5) [58, 64]. CATH's latest release introduced a new grouping (Class 6) called 'Special', containing non-globular structures including linker regions between domains, fragments, and short and synthetic peptides [65]. Class 6 also includes low-resolution structures. There are currently 790 Class 6 families.

- **Architecture / A-level:** Domains of the same class are subdivided into distinct A-level groups, based the spatial arrangement of secondary structural elements, but independent of the connectivity between secondary structure elements (see Figure 10) [58]).
- **Topology / T-level:** Members of same topology group share similar overall shape *and* connectivity of the secondary structure elements. Members of a topology group have similar structures, but may have diverse functions.
- **Homologous superfamily / H-level:** Homologous superfamily / H-level: Domains with significant structural similarity or sequence similarity and, usually, similar functions, putatively descended from a common ancestor.

CATH assigns to each domain a unique identifier, specifying its classification at different levels. For example, Sperm Whale myoglobin is 1.10.490.10:

Level	**CATH Code**	**Description**
Class	1	Mainly-α
Architecture	1.10	Orthogonal Bundle
Topology	1.10.490	Globin-like
Homologous Superfamily	1.10.490.10	Globins

figure_10a.eps here **figure_10b.eps here**

CATH clusters members of each homologous superfamily into structurally similar groups (SSGs). Structures within a homologous superfamily are superposed (clustered) at RMSD cutoffs of <5 Å and <9 Å to form tight and loose structural clusters. SSGs are helpful to understand the structural diversity of a superfamily. For each structure cluster, FunTree provides a phylogenetic tree and associated annotations ([67] `http://cpmb.lshtm.ac.uk/FunTree/`).

CATH also integrates additional functional annotations from UniProtKB such as Gene Ontology (GO), Enzyme Classification (EC), catalytic sites (from MACiE, CSA) and species information. More recently, CATH has sub-clustered members of Homologous superfamilies (H) into functionally coherent groups – Functional Families (FunFams): close relatives that share functions. (`http://cathdb.info/wiki/doku/?id=data:index`), see Figure 11. These contain the sequences of known structures and sequences of domains predicted to belong to the superfamily. FunFams can be used for predicting Gene Ontology (GO) [68] functions for uncharacterized sequences. The FunFam-based Function prediction pipeline has consistently performed well (among the top 5) for prediction of Gene Ontology Molecular Function and Biological Process terms in CAFA (Critical Assessment of Functional Annotation) [69]. The CATH website provides a sequence-based search utility for identifying FunFams using query protein sequences (`cathdb.info/search/by_sequence`), or through the Applications Program Interface (`https://github.com/UCLOrengoGroup/cath-api-docs`).

figure_11.jpg here

CATH also classifies protein domain *sequences* from UniProt that have been predicted to belong to specific CATH superfamilies even in the absence of experimentally-determined structures. This greatly expands the set of classified proteins, from 450,000 protein domains of the

PDB, to 150 million protein domains. The additional functional annotations adduced enable more comprehensive analysis of functional divergence within each superfamily.

4.2 SCOP and related databases (SCOPe, SCOP2)

A.G. Murzin and coworkers first developed the SCOP database in December 1994 [45]. They released the most recent updated version (1.75) in June, 2009 (`http://scop.mrc-lmb.cam.ac.uk/legacy/`). This version of SCOP is no longer being maintained and updated. Instead, SCOP has spawned:

1. A new project framework, SCOPe developed by Brenner and co-workers [50]. SCOPe (Structural Classification of Proteins — extended) is a development of SCOP version 1.75, and aims to update the original SCOP hierarchy.
2. SCOP2, developed by Murzin and co-workers, generalises the tree-like SCOP hierarchy.

4.2.1 Structural classification strategies

Classification protocol for SCOP: Domains are classified in SCOP by manual curation [45], although results from structure comparison algorithms can suggest or support relationships. SCOP classifies domains hierarchically on the basis of common structural and evolutionary relationships.
Classification protocol for SCOPe: Manual curation of superfamilies remains a key feature of SCOPe, in which very distant homologues with similar 3D structure and no recognizable sequence similarity are divided into homologues and possible analogues at the superfamily level, on the basis of the insight of human curators [71].
Classification protocol for SCOP2: The SCOP2 classification differs from that of simple tree-like hierarchy in SCOP [72–74]. In SCOP2, protein classification is described by a directed acyclic graph in which nodes form a network of many-to-many relationships. Here, the protein domain corresponds to a child node of the SCOP2 graph. Its boundaries are dependent on, and can *vary* with, each individual relationship. That is, each family and superfamily relationship in SCOP2 refers to a region of a protein sequence and structure; different relationships may link different, even overlapping, substructures of proteins. For example, Fold classification is an attribute of a single structural domain; however, a Family can span more than one structural domain within a protein.

4.2.2 The original SCOP had a tree-like classification

SCOP, and SCOPe, classify protein domains into a hierarchy of categories: Class, Fold, Superfamily, Family, Protein, and Species. Similar to CATH, Class and Fold (Topology) levels of classification rely on structural features and similarities. Levels Family and below are related by evolution; levels Superfamily and above indicate structural but not necessarily evolutionary relationship.

- **Class:** SCOP defines major classes:
    1. All-α proteins
    2. All-β proteins
    3. α/β proteins
    4. $\alpha + \beta$ proteins
    5. Multi-domain proteins

6. Membrane and cell surface proteins and peptides
7. Small proteins

Subsidiary classes are:

8. Coiled-coil proteins
9. Low resolution protein structures
10. Peptides
11. Designed (non-natural) proteins

The Class level in SCOP corresponds roughly to the Architecture level in CATH and ECOD.

- **Fold:** Domains comprising the same major secondary structures with the same spatial arrangement and topology of connection. The structural similarities *may* arise from preferred packing arrangements of secondary structure elements, and thus do not automatically imply evolutionary relationship.

  Colloquial architectural descriptions are given for folds showing generic arrangements of secondary structures, *e.g.,* annotation as sandwich or barrel structures.

- **Superfamily:** Proteins with enough similarity of structure and, often, function, to suggest an evolutionary relationship (evidence sufficient to indict but not necessarily to convict).

- **Family:** comprises genuinely-homologous proteins with similar sequences but typically with distinct functions. A query domain is assigned to a protein family if the % sequence identity to a previously classified domain is >55%, or if the gene annotation given in a query PDB structure matches the gene annotation of exactly one family in the assigned Superfamily.

- **Protein:** This level groups together similar domains that originate from different biological species or represent different isoforms – including paralogues – within the same species. This level is assigned if the % identity to the previously classified domain is ≥99%, or if the gene annotation provided in the query PDB structure matched the gene annotation of exactly one protein in the assigned Family.

- **Species:** Species represents a distinct protein sequence and its naturally-occurring or artificially-generated variants. (Not necessarily a biological species in the Linnaean sense.)

Referring to Figure 7, to compare with CATH, SCOP classifies and annotates the domains from 1ffh and 1rty as follows:

- **1ffh:**
  - **Fold:** Ferritin-like

    annotation: *core: 4 helices; bundle, closed or partly opened, left-handed twist; up-and-down*
  - **Superfamily:** Domain of the SRP/SRP receptor G-proteins
- **1rty:**
  - **Fold:** Ferritin-like

annotation: *core: 4 helices; bundle, closed, left-handed twist; 1 crossover connection*

– **Superfamily:** Cobalamin adenosyltransferase-like

annotation: *crossover loop goes across a different side of the 4-helical bundle; no internal metal-binding site*

4.2.3 SCOP2

SCOP2 uses a new classification approach in which protein relationships take the form of a directed acyclic graph, instead of a hierarchical tree. The nodes represent selected regions of proteins and relationships among them. Some of these are inherited from the original hierarchy (with some modifications). But more complex interrelationships abound, to allow more flexibility. The generalization proved necessary to describe complex evolutionary pathways and other subtle relationships. For example, domain rearrangements are examples of evolutionary events which are substantially more complex than sequence divergence within a domain by a succession of point mutations that maintain a common folding pattern.

To appreciate the challenges that required generalising SCOP2 from SCOP, consider the enzymes pyruvate decarboxylase and transketolase. These present a fairly complex example of retention of structural similarity and domain rearrangement, producing functional change (see Figure 12). The reader will readily recognise that the relationship between these two proteins is not describable by a strictly hierarchical scheme involving only the domains separately.

(a) **(b)** **(c)**

figure_12a.eps here **figure_12b.eps here** **figure_12c.eps here**

4.2.4 Classification in SCOP2

The highest level of classification in SCOP2 is the Protein type: soluble, membrane, fibrous, and intrinsically disordered. Within these are Classes:

1.	all-α	secondary structure predominantly α-helical
2.	all-β	secondary structure predominantly β-sheet
3.	α/β	containing both helices and sheets, separated in the structure
4.	$\alpha+\beta$	containing alternating $\alpha-$helices and strands of β-sheet
5.	small proteins	containing little or no secondary structure

Proteins within each class are assigned to Fold groups, on the basis of global architectural features, including the secondary structure content and the topology of assembly. Similarity of overall architecture does not necessarily imply evolutionary relationship.

Fold groups are divided into Superfamilies, and subdivided into Families, on the basis of structural relationship. The Family and Superfamily are the basic evolutionary levels in SCOP2. A Family is a group of closely-related proteins with clear evidence of homology. Members of a family may be multi-domain proteins. A Superfamily is a grouping of single-domain components

of families, comprising proteins that probably originated in a common ancestor, but without the intimacy of relationship that would link them into the same family. Some superfamilies are classified as Intrinsically Unstructured Protein Regions (IUPRs). These exhibit multiple conformations, or are even entirely unstructured in the free state, although they may adopt a structured conformation – or even several different structured conformations – in different states of ligation.

Many proteins contain multiple domains unrelated in structure to one another. Figure 13 shows an example of a protein in the GH64 $\beta-$glucanase-like family, containing two domains from different superfamilies. Note that the subgraph shown under **Show ancestry** is NOT a tree with the highest level, class, as its root.[†]

figure_13.jpg

Protein relationships in SCOP2: Protein relationships are categorized as Evolutionary and Structural.

- *Evolutionary Relationships* include SCOP-like levels: Family and Superfamily; see, for example, the **Show ancestry** section in Figure 13). SCOP2 introduced an additional level above of Superfamily – namely 'Hyperfamily' (see Table 4).

Table 4 here

- *Structural relationships* in SCOP2 are treated as a separate category from evolutionary relationships. This allows more flexible classification of evolutionarily-related but structurally distinct proteins.

SCOP2 entries are also very rich in annotation, and in links to other databases and software. Whereas annotation in SCOP was, in effect, a collection of 'marginal notes'; in SCOP2 annotations are more formal, observing in most cases a controlled vocabulary (keywords and tags). Conversely, several SCOP2-based groupings/relationships are included as structural annotations in the more recent SCOP classifications.

SCOP will be continued as part of the 3D-SCAfold project being developed at PDBe, with initial funding from Wellcome Trust and MRC-LMB. This will establish an automated platform for domain boundary identification and homologue recognition for the SCOP and CATH databases. Very remote homologues will be validated by manual curation.

4.3 The ECOD database

ECOD (Evolutionary Classification of protein Domains) is a hierarchy comprising five levels: architecture (A), possible homology (X), homology (H), topology (T), and family (F). In grouping protein domains, ECOD emphasises evolutionary relationships rather than folding pattern.

[†]In this example, the subgraph shown *could* be regarded as a tree, starting with the family node as the root. (Indeed, having the root at the bottom would more accurately reflect the botanical metaphor.) However, in many cases the ancestry subgraph is a directed acyclic graph but not a tree.

- Architecture: As in CATH, the architecture level (A) groups domains with similar secondary structure compositions and geometric shape.
- Possible homology (or X-level): This level groups domains for which homology is suggested – usually by overall structural (or fold) similarity – but not provable.
- Homology (or H-level): This groups domains that are descended from a common ancestor as shown by significant sequence and/or structure scores, shared functional properties, opinions in literature and in SCOP, *etc.*
- Topology (or T-level): T-level occurs *under* H-level, with the insight that homologues can have different topologies [77]. Within the same homologous superfamily, or H group, ECOD classifies Homologues with distinct topologies into different T-groups.
- Family (or F-level): This level groups domains that have significant sequence similarity, primarily based on Pfam.

The unique feature of the ECOD classification scheme is that within the homology H-level it further classifies domains by fold or topology, *i.e.,* T-groups. This accommodates the significant structural changes that can occur between very remote homologues. This level can provide invaluable insights into the mechanisms by which domain structures within a particular superfamily change during evolution. Indeed, compared to other structural classification resources, ECOD catalogues a very high number of evolutionary links among classified structural domains, by accounting for extremely distantly-related homologs.

5 Other general macromolecular structure databases

Several databases present related data:

- The Nucleic Acid Database (NDB) collects information about experimentally-determined nucleic acids and complex assemblies (`http://ndbserver.rutgers.edu`). It overlaps in contents with the PDB.
- The Cambridge Structural Database, based at the Cambridge Crystallographic Data Centre (CCDC), presents structural data on small-molecule organic and metal-organic compounds (`www.ccdc.cam.ac.uk`).
- The wwPDB has launched an archive, PDB-Dev, for structural models and the underlying data obtained through use of integrative methods utilising a combination of complementary experimental and computational techniques (`http://pdb-dev.wwpdb.org`).

Many other projects present data from the PDB, but reorganise or re-present it, often contributing different sets of added annotations. These include:

Jena Library of Biological Macromolecules	Annotatory information about PDB structures; emphasis on visualization and analysis
Molecular Modelling DataBase (MMDB)	NCBI's ENTREZ 3-D structure database. Contains experimentally-determined structures from PDB – but, despite its name – *not* predicted models.
OCA (Goose in a number of Romance languages, also the letters precede PDB in the alphabet)	a browser-database with emphasis on sequence-structure-function relationships
PDBsum	A pictorial database providing an overview of each structure deposited in the PDB.
PDBWiki	Community-annotated knowledge base of biological molecular structures and related information.
Proteopedia	Encyclopedia of macromolecular structures. Proteopedia contains a page for every PDB entry, plus user-provided interactive pages that are aimed at presenting structure/function information to a broad scientific audience.
SFLD	The Structure-Function Linkage Database (SFLD) describes protein evolution within families as a network, with nodes labelled by function and linked to the PDB [78].

For a more complete list, including URLs, see `https://bip.weizmann.ac.il/toolbox/structure/databases.html`.

5.1 'Boutique' databases

These contain information of interest to specialists in particular classes of proteins; for instance, meeting announcements. (See: `https://bip.weizmann.ac.il/toolbox/structure/databases.html`)

A few popular examples:

Topic	URL
Antibodies	`http://www.bioinf.org.uk/abs/`
Proteases	`https://www.ebi.ac.uk/merops/`
G-protein-coupled receptors	`https://gpcrdb.org/`
Membrane proteins	`https://blanco.biomol.uci.edu/MemPro_resources.html`

6 Sequences and structures

Because of the success of high-throughput DNA sequencing techniques, *a great deal more* amino-acid sequence data are available, than three-dimensional protein structural data. A convenient and reliable method for predicting protein structure from amino-acid sequence could eliminate this disparity. Indeed, recent Critical Assessment of Structure Prediction (CASP) programs have shown very spectacular breakthroughs.

There are two main types of successful structure prediction: comparative, or homology, modelling; and *a priori* structure prediction;

• *Comparative modelling* is the prediction of an unknown structure, knowing the structures of one or more related proteins.
• a priori *structure prediction* is the prediction of an unknown structure, without making *explicit* use of any other known structure. (However, many algorithms make use of *general* properties of known structures; for example distributions of observed bond lengths and angles, or they may use small fragments of known structures, or may use machine-learning methods involving training on a large set of known structures.)

6.1 Comparative modelling

In homology modelling, the first step is to identify one or more close relatives of the target protein, of known structure. Under the assumption that similarity of sequence implies similarity of structure [51], the next step is to construct a model of the target protein from the structures of the relatives: What are the *differences* between the structure of the target protein and the structures of the relatives? (If *a priori* structure prediction is the *integral* form of the structure prediction problem, homology modelling is the *differential* form.)

There are several mature homology modelling programs, which allow users to submit novel sequences to associated web servers (See Table 5). I-TASSER, by Y. Zhang and colleagues, has been very successful in recent CASP programs. The 3D-Beacons project will combine experimentally-determined and predicted structures (`https://gtr.ukri.org/projects?ref=BB%2FS020144%2F1`).

Modelling program	Lead author	Web server
I-TASSER	Y. Zhang	`https://zhanglab.ccmb.med.umich.edu/I-TASSER/`
SWISS-MODEL	T. Schwede	`https://swissmodel.expasy.org/`
MODELLER	A. Sali	`https://salilab.org/modeller/`
Phyre[2]	M.J.E. Sternberg	`http://www.sbg.bio.ic.ac.uk/phyre2`
PSIPRED	D. Jones	`http://bioinf.cs.ucl.ac.uk/web_servers/psipred_server/psipred_overview/`

Table 5. Homology-modelling servers

New amino-acid sequences are regularly run through homology-modelling programs, and the accumulated results made available in repositories:

Repository	**URL**
SWISS-MODEL repository	`http://swissmodel.expasy.org/repository`
ModBase	`https://modbase.compbio.ucsf.edu/modbase-cgi/index.cgi`
Biological Structure Model Archive (BSMA)	`https://bsma.pdbj.org/`
ModelArchive	`https://modelarchive.org/`
Genome3D	`http://genome3d.eu/`

6.2 *a priori* prediction

Recently, methods of structure prediction *not* based on specific known relatives of the target protein have achieved major breakthroughs.

In the latest completed CASP programme (CASP-14, 2020), a program AlphaFold2 by Alphabet subsidiary DeepMind showed spectacular success [79]. The approach made use of tech-

niques of artificial intelligence, a field made famous by computer programs able to beat the human world champions in Chess, and Go.

There is consensus that this breakthrough is a real 'game-changer'. It will revolutionise the relationship between sequence and structure databases. This achievement has been the subject of a spate of papers in Nature, Science and other journals, including a special dedicated issue of The Journal of Molecular Biology.

Figure 14 shows the predictions by AlphaFold2 of two targets from CASP14 (2020). The chain tracings of the experimental result and the prediction are virtually identical. These two examples are not selected rare successes; AlphaFold2 achieved comparable results consistently. Nearly two-thirds of its predictions are comparable in quality to experimentally-determined structures. Indeed, John Moult, the originator and leader of the CASP programs since its inception in 1994 – and who must deservedly be dead chuffed at the triumph of the effort – commented that: 'In some cases, it was not clear whether the discrepancy between AlphaFold's predictions and the experimental results was a prediction error or an artefact of the experiment.'

figure_14a.jpg here **figure_14b.jpg here**

Although protein structure prediction programs have achieved very impressive, and even fairly consistent successes it would be wise to remember that the results are NOT supported directly by experimental data. However, it should be noted that AlphaFold2 provides reliable residue-by-residue estimates of confidence in its predictions.

DeepMind, the authors of AlphaFold2, and EMBL's European Bioinformatics Institute (EMBL-EBI) are partners in developing a database, AlphaFold DB, to make these results freely available to the scientific community. The current release of this database covers the human proteome and the proteomes of several other key organisms. 365,000 predicted structures are available now, and it is expected that the database will grow to 100,000,000 entries before the end of 2021 (see https://alphafold.ebi.ac.uk/).

7 Protein interactions

In order to understand the activities of proteins within cells, we need to know the sets of interacting molecules. These include protein-protein and protein-nucleic acid interactions, as well of course as interactions with cofactors and metabolites. There are three basic sources of information:

1. Structure determination. Results include PDB entries containing multimers or complexes, revealing not only the partners but the detailed mode of interaction. Cryo-EM is making it easier to determine the structures of large assemblies. Databases include (the first two are both based at the European Institute of Bioinformatics; although independent of the PDBe they share data):

IntAct	`https://www.ebi.ac.uk/intact/`
Complex Portal	`https://www.ebi.ac.uk/complexportal/home`
String	`https://string-db.org/cgi/input.pl`
Interactome3D	`https://interactome3d.irbbarcelona.org/`

   Other databases explicitly collect interfaces:

PIBASE	A database of structurally defined protein interfaces `https://modbase.compbio.ucsf.edu/pibase/queries.html`
SNAPPI	Structures, interfaces and alignments for protein-protein interactions `http://www.compbio.dundee.ac.uk/SNAPPI/`
3DID	3D interacting domains `https://3did.irbbarcelona.org/`

2. Experimental methods that detect interactions, but do not determine structures, include:

   - Two-hybrid screening systems
   - Chemical cross-linking
   - Coimmunoprecipitation
   - Chromatin immunoprecipitation
   - Phage display
   - Mass spectrometry
   - Surface plasmon resonance
   - Fluorescence resonance energy transfer (FRET)
   - Tandem affinity purification
   - Domain recombination networks (a computational method)
   - Coexpression patterns
   - Phylogenetic distribution patterns
   - Sequence co-evolution is a computational method for identifying protein complexes and the contacts between components [80].

3. Analysis of the literature. Some projects 'manually' harvest data from publications. Others automate the process by computer processing of texts of articles, searching the scientific literature for sentences of the form:

   <protein name> ... binds ... <protein name>

   and collect the indicated partners. (This is obviously a very oversimplified description of the actual challenge.)

   Several databases collect interaction patterns:

BioGRID	The biological general repository for interaction datasets
DIP	Database of interacting proteins
DOMINE	Database of protein domain interaction
HPID	Human protein interaction database
HPRD	The human protein reference database
I2D	Interolog interaction database
iHOP	Protein association network built by literature mining in PubMed
MINT	Molecular interaction database
MIPS	Mammalian protein-protein interaction database
UniHI	Unified human interactome

A list appears at:
http://www.vls3d.com/index.php/links/bioinformatics/protein-protein-interaction/ppi-databases-network

In addition to these experimental methods, there is the possibility of taking known structures of individual protein components of complexes, and computationally predicting their modes of association. This is known as 'the docking problem'. CAPRI (Critical Assessment of PRediction of Interactions https://www.ebi.ac.uk/msd-srv/capri/) runs blind tests, like those organised by CASP, to evaluate docking methods [81, 82].

What makes the docking problem difficult are the conformational changes generally observed, between the structures of the components in isolation and in the complex.

A related problem, important in applications to drug design, is the prediction of interactions and affinities of small molecules to proteins. Given the possibility of considering larger molecules, including peptides, as drugs, these problems overlap.

Although web sites offering docking calculations exist, at the time of writing the authors are not aware of any algorithms as successful as those that produce *a priori* predictions of structures of individual domains from amino-acid sequences. A special case is the building of homology models of complexes from known structures of related complexes.

A more general docking problem is the prediction of the structure of a complex containing *more* than two proteins from known structures of the components. However, we are not aware of any successful programs that address this problem. Of course one could attempt pairwise docking of components of the complex, but that is not a general approach. (If the multiple-docking problem were solved effectively, one could envisage testing all combinations of PDB entries for potential interactions, but this is currently not in sight.)

8 Expected developments

The reader will understand that the following suggestions are associated with varying degrees of confidence!

1. There will undoubtedly be a continued and almost certainly an accelerated growth in the PDB, both in terms of numbers of entries and numbers of residues. Richard Henderson has predicted that in a few years the rate or production of electron cryo-microscopy structures will exceed that of X-ray crystal structures.

2. It is very likely that results of structure prediction methods will improve to the point where it will be possible to convert large-scale sequence databanks into adjunct structure databanks. There are two avenues for this: A threshold will be reached when the PDB is sufficiently comprehensive for everything else to be built by homology modelling. (This was one projected outcome of structural genomics projects [83].) The second is the burgeoning success of *a priori* structure prediction.

3. One extension of these developments will be to modelling of protein dynamics – conformational changes during activity of individual proteins (*e.g.,* [84, 85]), and possibly even simulation of protein folding pathways.

4. Another extension will be continued development of methods for the creation of proteins with properties not known in nature – either by directed evolution [86] or *a priori* design [87, 88],

- For instance, faced with a pandemic, knowing the structure of the SARS-Cov-2 spike protein, design by a computer program a set of minimally-antigenic antibodies that will bind it with high affinity ... then synthesise the corresponding genes, insert into *E. coli,* and harvest and distribute the antibodies ...

5. A desired development is better integration of interactions, including growth of structure determinations of protein and protein-nucleic acid complexes – supported by advances in electron cryo-microscopy – and supplemented by successful docking software (for pairwise and oligomeric complexes, although one would be justified in being more pessimistic about imminent progress towards this goal). This will achieve better understanding of, and ability to simulate, regulatory processes, many of which are mediated by protein-protein and protein-DNA interactions. (An ambitious project, prominently including but not limited to analysis at the level of molecular structure, is described at `https://www.pbcconsortium.org/`.) Will it someday be possible to model an entire cell at molecular resolution?[‡]

9 Conclusions

Dare we ask: What would it take for 'wet labs' to disappear from academic biology? (Clearly sequencing will be essential in clinical applications; for example in tracing evolution of viral strains.) If we knew all the sequences and structures and functions of all proteins and nucleic acids of all living and many extinct organisms, could biology – at the molecular level at least – become a topic within computer science, whereby we answer all questions by data retrieval and simulation?

Personally, we hope not; but we are afraid that we can't give any guarantees.[§]

Acknowledgements

We thank A.G. Murzin and A. Andreeva for helpful advice.

[‡]Sydney Brenner used to chaff Aaron Klug, saying: 'Why don't you crystallise *E. coli?*

[§]Many readers will recall that Paul Dirac famously made a similar-sounding claim about chemistry, in 1929, but this has not happened.

References

[1] Anonymous (1971). Crystallography: Protein Data Bank. Nature New Biol. 233, 223.

[2] Anonymous (2021). A celebration of structural biology. Nat. Methods 18, 427.

[3] Bordin, N., Sillitoe, I., Lees, J.G. & Orengo, C. (2021). Tracing evolution through protein structures: Nature captured in a few thousand folds. Front. Mol. Biosci. 8, 408.

[4] Dayhoff, M.O., Eck, R.V. *et al.* (1965). Atlas of Protein Sequence and Structure. Silver Spring, Maryland: National Biomedical Research Foundation.

[5] Lipscomb, W.N., Reeke G.N., Jr, Hartsuck, J.A., Quiocho, F.A. & Bethge, P.H. (1970). The structure of carboxypeptidase A. 8. Atomic interpretation at 0.2 nm resolution, a new study of the complex of glycyl-L-tyrosine with CPA, and mechanistic deductions. Phil. Trans. Roy. Soc. Lond. B257, 177–214.

[6] Berman, H., Henrick, K. & Nakamura, H. (2003). Announcing the worldwide Protein Data Bank Nature Struct. Mol. Biol. 10, 980.

[7] wwPDB consortium. (2019). Protein Data Bank: the single global archive for 3D macromolecular structure data. Nucl. Acids Res. 47, D520–D528.

[8] Lesk, A.M. (2016). Introduction to Protein Science, 3rd. ed. (Oxford: Oxford University Press)

[9] Seoane, B. & Carbone, A. (2021). The complexity of protein interactions unravelled from structural disorder. PLoS Comput. Biol. 17:e1008546.

[10] Borrell, B. (2009). Fraud rocks protein community. Nature 462, 970.

[11] Young, J.Y. *et al.* (2017). OneDep: Unified wwPDB system for deposition, biocuration, and validation of macromolecular structures in the PDB archive. Structure 25, 536–545.

[12] Baker, E.N. & Saenger, W. (1999). Deposition and release of macromolecular structural data. Acta Cryst. D55, 2.

[13] Joosten R.P. & Vriend G. (2007). PDB improvement starts with data deposition. Science 317,195–196.

[14] Commission on Biological Macromolecules (2000). Guidelines for the deposition and release of macromolecular coordinate and experimental data. Acta Cryst. D56, 2.

[15] Gore, S., Sanz-Garcia, E., Hendrickx, P.M.S., Gutmanas, A., Westbook, J.D., *et al.* (2017). Validation of structures in the protein data bank. Structure 25, 1916-1927.

[16] Berjanskii, M., Zhou J., Liang Y., Li G. & Wishart, D.S. (2012). Resolution-by-proxy: a simple measure for assessing and comparing the overall quality of NMR protein structures. J. Biomol. NMR. 53, 167–180

[17] Lawson, C.L. & Chiu, W. (2018). Comparing Cryo-EM structures. J. Struct. Biol. 204, 523–526.

[18] Lawson, C.L., Berman, H.M. & Chiu, W. (2020). Evolving data standards for cryo-EM structures. Structural Dynamics 7, 014701.

[19] Lange, J., Baakman, C., Pistorius, A., Krieger, E., Hooft, R., Joosten, R.P. & Vriend, G. (2020). Facilities that make the PDB data collection more powerful. Protein Sci. 29, 330–344.

[20] Joosten, R.P., Womack, T., Vriend, G. & Bricogne, G. (2009). Re-refinement from deposited X-ray data can deliver improved models for most PDB entries. Acta Cryst. D65, 176–185.

[21] Joosten, R.P. *et al.* (2009) PDB_REDO: automated re-refinement of X-ray structure models in the PDB. J. Appl. Cryst. 42,376–384.

[22] Joosten, R.P., Joosten, K., Cohen, S.X., Vriend, G. & Perrakis, A. (2011). Automatic rebuilding and optimization of crystallographic structures in the Protein Data Bank. Bioinformatics 27, 3392–3398.

[23] Wilkinson, M.D., Dumontier, M., Aalbersberg, IJ.J., Appelton, G., Axton, M. *et al.* (2016). The FAIR Guiding Principles for scientific datamanagement and stewardship. Scientific Data 3, 160018.

[24] Armstrong, D.R., Berrisford, J.M., Conroy, M.J., Gutmanas, A., Anyango, S. *et al.* (2020). PDBe: improved findability of macromolecular structure data in the PDB. Nucl. Acids Res. 48, D335–D343.

[25] Coker, E.A. *et al.* (2019). canSAR: update to the cancer translational research and drug discovery knowledgebase. Nucl. Acids Res. 47, D917–D922.

[26] Orengo, C. *et al.,* 2020. A community proposal to integrate structural bioinformatics activities in ELIXIR (3D-Bioinfo Community) F1000Res. 9, ELIXIR-278.

[27] de Chadarevian, S. (2018). John Kendrew and myoglobin: Protein structure determination in the 1950s. Prot. Sci. 27, 1136–1143.

[28] Phillips, S.E. (2018). Structure and refinement of oxymyoglobin at 1.6 Å resolution. J. Mol. Biol. 142, 531-554.

[29] Altschul, S.F., Gish, W., Miller, W., Meyers, E.W. & Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

[30] Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z, Miller, W. & Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl. Acids Res. 25, 3389-3402.

[31] Krogh, A., Brown, B., Mian, I.S., Sjolander, K. & Haussler, D. (1994). Hidden Markov models in computational biology: applications to protein modeling. J. Mol. Bio. 235, 1501–1531.

[32] Eddy, S.R. (1996). Hidden Markov models. Curr. Opin. Struct. Biol. 6, 361–365.

[33] Eddy, S.R. (1998). Profile Hidden Markov Models. Bioinformatics, 14, 755–763.

[34] Mirdita, M., Steinegger, M. & Söding, J. (2019). MMseqs2 desktop and local web server app for fast, interactive sequence searches. Bioinformatics 35, 2856–2858.

[35] Krissinel, E. & Henrick, K. (2004). Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Cryst. D60, 2256–2268.

[36] Burley, S.K., Bhikadiya, C., Bi, C., Bittrich, S., Chen, L. *et al.* (2021). RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. Nucl. Acids Res. 49, D437–D451.

[37] Sumanaweera, D., Allison, L, & Konagurthu, A.S. (2019). Statistical compression of protein sequences and inference of marginal probability landscapes over competing alignments using finite state models and Dirichlet priors. Bioinformatics. 35, i360–i369.

[38] Collier, J.H., Allison, L., Lesk, A.M., Stuckey, P.J., Gardia de la Banda, M. & Konagurthu, A.S. (2017). Statistical inference of protein structural alignments using information and compression, Bioinformatics. 33, 1005–1013.

[39] Konagurthu, A.S., Whisstock, J.C., Stuckey, P.J., Lesk, A.M. (2006). MUSTANG: a multiple structural alignment algorithm. Proteins: Structure, Function, Bioinformatics 64, 559–574.

[40] Collier, J.H., Lesk, A.M., Garcia de la Banda, M. & Konagurthu, A.S. (2012). Super: a web server to rapidly screen superposable oligopeptide fragments from the protein data bank. Nucl. Acids Res. 40, W334–W339.

[41] Konagurthu, A.S., Lesk, A.M. & Allison, L. (2012). Minimum Message Length inference of secondary structure from protein coordinate data. Bioinformatics 28, i97–i105.

[42] Konagurthu, A.S., Subramanian, R., Allison, L, Abramson, D., Stuckey, P.J., Gardia de la Banda, M. & Lesk, A.M. (2021). Universal architectural concepts underlying protein folding patterns. Frontiers in Molecular Biosciences 7, 491.

[43] Bourne, P.E., Berman, H.M., McMahon, B. Watenpaugh, K.D., Westbrook, J. & Fitzgerald, P.M.D. (1977). The Macromolecular Crystallographic Information File (mmCIF). Methods in Enzymology 277, 571–590.

[44] Wetlaufer, D.B. (1973). Nucleation, rapid folding, and globular intrachain regions in proteins. Proc. Nat'l. Acad. Sci. USA. 70, 697–701.

[45] Murzin, A.G., Brenner, S.E., 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.

[46] Orengo, C., Jones, D. & Thornton, J. (1994). Protein superfamilies and domain superfolds. Nature 372, 631–634.

[47] Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B. & Thornton, J.M. (1997). CATH – a hierarchic classification of protein domain structures. Structure 5, 1093–1108.

[48] Chandonia, J.M. & Brenner, S.E. (2006). The impact of structural genomics: expectations and outcomes. Science 311, 347–351.

[49] Fox, N.K., Brenner, S.E. & Chandonia, J.M. (2015). The value of protein structure classification information—Surveying the scientific literature. Proteins 83, 2025–2038.

[50] Fox, N.K., Brenner, S.E. & Chandonia, J.M. (2014). SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic Acids Res. 42, D304–309.

[51] Lesk, A.M. & Chothia, C. (1986). The response of protein structures to amino acid sequence changes. Phil. Trans. Roy. Soc. London, A317, 345–356.

[52] Greene, L.H., Lewis, T.E, Addou, S., Cuff, A., Dallman, T. *et al.* (2007). The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. Nucleic Acids Res. 35, D291–D297.

[53] Taylor W.R. & Orengo, C.A. (1989). Protein structure alignment. J. Mol. Biol. 208, 1–22

[54] Orengo, C.A. & Taylor, W.R. (1996). SSAP: Sequential structure alignment program for protein structure comparison. Meth. Enzymol. 266, 617–635.

[55] Cheng, H., Liao, Y., Schaeffer, R.D. & Grishin, N.V. (2015). Manual classification strategies in the ECOD database. Proteins 83, 1238–1251.

[56] Cheng. H., Schaeffer, R.D., Liao, Y. *et al.* (2014). ECOD: an evolutionary classification of protein domains. PLoS Comput Biol. 10:e1003926.

[57] Orengo, C., Jones, D. & Thornton, J. (1994). Protein superfamilies and domain superfolds. Nature 372, 631–634.

[58] Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B. & Thornton, J.M. (1997). CATH – a hierarchic classification of protein domain structures. Structure 5, 1093–1108.

[59] Sillitoe, I., Dawson, N., Thornton, J. & Orengo, C. (2015). The history of the CATH structural classification of protein domains. Biochemie 119, 209–217.

[60] Sillitoe, I., Dawson, N., Lewis, T.E., Das, S., Lees, J.G. *et al.* (2019). CATH: expanding the horizons of structure-based functional annotations for genome sequences. Nucleic Acids Res. 47, D280–D284.

[61] Sillitoe, I., Bordin, N., Dawson, N., Waman, V.P. *et al.* (2021). CATH: increased structural coverage of functional space. Nucleic Acids Res. 49, D266–D273.

[62] Das, S., Sillitoe, I., Lee, D., Lees, J.G., Dawson, N.L., Ward, J. & Orengo, C.A. (2015). CATH FunFHMMer web server: protein functional annotations using functional family assignments. Nucleic Acids Res. 43, W148–153.

[63] Levitt, M. & Chothia, C. (1976). Structural patterns in globular proteins. Nature 261, 552–558.

[64] Michie, A.D., Orengo, C.A., Thornton, J.M. (1996). Analysis of domain structural class using an automated class assignment protocol. J. Mol. Biol. 262, 168–185.

[65] Sillitoe, I., Bordin, N., Dawson, N., Waman, V.P., Ashford, P. *et al.,* (2021). CATH: increased structural coverage of functional space. Nucl. Acids. Res. 49, D226-273.

[66] Presnell, S.R. & Cohen, F.E. (1989). Topological distribution of four-α-helix bundles. Proc. Nat'l. Acad. Sci. USA. 86, 6592–6596.

[67] Furnham, N., Sillitoe, I., Holliday, G.L. *et al.* (2012) FunTree: a resource for exploring the functional evolution of structurally defined enzyme superfamilies. Nucleic Acids Res. 40, D776–D782.

[68] The Gene Ontology Consortium: Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H. *et al.* (2000). Gene ontology: tool for the unification of biology. Nat. Genet. 25, 25-29.

[69] Zhou, N., Jiang, Y., Bergquist, T.R., *et al.* (2019). The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. Genome Biol. 20, 244.

[70] Valdar, W.S. & Thornton, J.M. (2001). Protein-protein interfaces: analysis of amino acid conservation in homodimers. Proteins 42, 108–124.

[71] Chandonia, J.M., Fox, N.K. & Brenner, S.E. (2017). SCOPe: Manual curation and artifact removal in the Structural Classification of Proteins – extended database. J. Mol. Biol. 429, 348–355.

[72] Andreeva, A., Howorth, D., Chothia, C., Kulesha, E. & Murzin, A.G. (2014). SCOP2 prototype: a new approach to protein structure mining, Nucleic Acids Research, 42, D310–D314.

[73] Andreeva, A., Howorth, D., Chothia, C., Kulesha, E. & Murzin, A.G. (2018). Investigating protein structure and evolution with SCOP2. Current Protocols in Bioinformatics 49, 1.26.1–1.26.21

[74] Andreeva, A., Kulesha, E., Gough, J. & Murzin, A.G. (2020). The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. Nucleic Acids Research 48, D376–D382.

[75] Lesk, A.M. (2021). Protein Science. (Oxford: Oxford University Press)

[76] Das, S., Dawson, N.L. & Orengo, C.A. (2015). Diversity in protein domain superfamilies. Curr. Opin. Genet. Dev. 35, 40-–49.

[77] Grishin, N.V. (2001). Fold change in evolution of protein structures. J. Struct. Biol. 134, 167–185.

[78] Akiva, E., Brown, S., Almonacid, D.E., Barber, A.E, II, Custer, A.F. *et al.* (2014). The Structure-Function Linkage Database. Nucl. Acids Res. 42, D521-530.

[79] Lupas, A.N., Pereira, J., Alva, V., Merino, F., Coles, M. & Hartmann, M.D. (2021). The breakthrough in protein structure prediction. Biochem. J. 478, 1885–1890.

[80] Hopf, T.A. *et al.* (2014). Sequence co-evolution gives 3D contacts and structures of protein complexes. eLife 3, e03430

[81] Lensink, M.F. *et al.* (2019) Blind prediction of homo- and hetero-protein complexes: The CASP13-CAPRI experiment. Proteins: Structure, Function and Bioinformatics 87, 1200–1221.

[82] Wodak, S., Velankar, S. & Sternberg, M.J.E. (2020). Modeling protein interactions and complexes in CAPRI: Seventh CAPRI evaluation meeting, April 3-5 EMBL-EBI, Hinxton, UK. Proteins: Structure, Function, Bioinformatics, 88, 911–912. (And other articles in that issue.)

[83] Vitkup, D., Melamud, E., Moult, J. & Sander, C. (2001). Completeness in structural genomics. Nature Struct. Biol. 8, 559–566.

[84] Gao, Y.Q., Yang, W. & Karplus, M. (2005). A structure-based model for the synthesis and hydrolysis of ATP by F1-ATPase. Cell 123, 195–205.

[85] Pu, J. & Karplus, M. (2008). How subunit coupling produces the γ-subunit rotary motion in F1-ATPase. Proc. Nat'l. Acad. Sci. USA 105, 1191–1197.

[86] Arnold, F.H. (2019). Innovation by evolution: Bringing new chemistry to life. Angew. Chem. Int. Ed. 58, 14420–14426

[87] Siegel, J.B., Zanghellini, A., Lovick, H.M., Kiss, G., Lambert, A.R., St Clair, J.L., Gallaher, J.L., Hilvert, D., Gelb, M.H., Stoddard, B.L., Houk, K.N., Michael, F.E. & Baker, D. (2010). Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction. Science. 329, 309–313.

[88] Privett, H.K., Kiss, G., Lee, T.M., Blomberg, R., Chica, R.A., Thomas, L.M., Hilvert, D., Houk, K.N. & Mayo, S.L. (2012). Iterative approach to computational enzyme design. Proc. Nat'l. Acad. Sci. USA. 109, 3790–3795.

Tables

Deposition type	Mandatory file uploads	Optional file uploads
X-ray and neutron crystallography	Atomic coordinates Structure factor data	Ligand definition file or image Auxiliary files, including raw data
Solution and solid-state NMR	Atomic coordinates Assigned chemical shifts Restraints used in refinement Auxiliary sequence file from AMBER	Spectral peak lists Ligand definition file or image Auxiliary files
Electron crystallography	Atomic coordinates Structure factor data or Electron potential map	Ligand definition file or image Auxiliary files
3D Electron microscopy (map and model)	Atomic coordinates Electric potential map Entry image for public display (EMDB)	Any number of additional maps Any number of masks Two half maps Fourier shell correlation (FSC) curve Ligand definition file or image
3D Electron Microscopy (map only)	Electric potential map Entry image for public display	As above
3D Electron Tomography	Tomogram Entry image for public display	As above

Table 1. Required deposition material (From [15].)

Database	**Group**	**Developed in**	**Website**
SCOP v1.75	Murzin	1994	`http://scop.mrc-lmb.cam.ac.uk/legacy/` (no longer maintained and updated)
SCOPe	Brenner / Chandonia	2014	`https://scop.berkeley.edu/`
SCOP2 (SCOP prototype)	Murzin	2014	`http://scop2.mrc-lmb.cam.ac.uk/`
CATH	Orengo	1994	`http://www.cathdb.info/`
ECOD	Grishin	2014	`http://prodata.swmed.edu/ecod/`

Table 2. Domain-based structural classification resources.

SCOP v1.75 Latest release June 2009	**SCOPe** Latest release January 2021	**SCOP2** Latest release January 2021	**CATH v4.3** Latest release November 2020	**ECOD** Latest release February 2021
Class (7)	Class (7)	Class (5)	Class (5)	–
–	–	–	Architecture (41)	Architecture (20)
Fold (1195) *Based on structural similarity*	Fold (1232) *Based on structural similarity*	Fold (1487) –	Topology/fold (1390) *Based on structural similarity*	X-group (2460) *Based on structural similarity*
–	–	IUPR (22)	–	–
–	–	Hyperfamily (18)	–	–
Superfamily (1962)	Superfamily (2026)	Superfamily (2660)	Homologous superfamily (6631)	H-Group (3715)
–	–	–	–	T-Group (∼3950)
Family (structures only) (3902)	Family (4919)	Family (5563)	Functional families (32,388)	F-level (16,300) *Based on structural similarity*
Domains (110,800)	Domains (325,245)	Domains (66,524)	Domains (500,238)	Domains (813,538)

Table 3. Structural Classification schemes in SCOP v1.75, SCOPe, SCOP2, CATH, and ECOD. The numbers in parentheses in the table entries report the populations of each of the categories.

Evolutionary relationships in SCOP2 [72–74]	
Level	**Definitions**
Hyperfamily	Hyperfamily represents the most structurally diverse SCOP superfamilies. Proteins in the same hyperfamily show some similarities in folding pattern but with some rearrangements of the topology suggesting *possible* distant homology.
Superfamily	Superfamily is represented by the common structural region shared by different protein families.
Family	Family corresponds to the conserved sequence region shared by closely related proteins.

Table 4. The different levels under Evolutionary Relationships, in SCOP2.

Figure captions

Figure 1. Growth in number of entries in PDB. Blue: Cumulative total number of entries. Red: Number of residues released in single calendar years. The red segment on the right is the number of entries deposited that year, on the same scale as the cumulative total.

Figure 2. Number of structures deposited per year, solved by different experimental methods: X-ray crystallography, NMR spectroscopy, and electron cryo-microscopy, and multiple methods, on logarithmic scale. Note the recent sharp increase in rate of production of electron cryo-microscopy structures. The category Multiple methods includes structures solved by two or more different experimental methods, for instance X-ray crystallography and NMR spectroscopy. From `https://www.rcsb.org/stats/all-released-structures`

Figure 3. Histogram of number of residues in PDB entries. From `http://www.rcsb.org/stats/distribution-residue-count`

Figure 4. (a) An illustration of the folding pattern of the backbone of acylphosphatase from the X-ray crystal structure. Chevrons indicate the strand direction (N→C). Strands of $\beta-$sheet depicted by large arrows; α-helices by translucent cylinders; loops between successive secondary structural elements by thin ribbons. (b) Five models of the backbone conformation, from the NMR structure determination. (c) The crystal structure (wide blue ribbon) superposed onto the NMR structures. [(a) from [8]]

Figure 5. The NCBD (IbID) domain of the creb-binding protein (CBP) is disordered in the unligated state. It forms different structures in complex with (left) the ACTR domain of p160 [1kbp] and (right) with IRF3 [1zoq]. In both pictures, the common NCBD domain is shown in blue. The C-terminal region of the NCBD domain, including two of the three helices, has approximately the same structure in both complexes. (From [8].)

Figure 6. Process of assimilation into the PDB of a new submitted entry; the OneDep system. Curation, including validation, acceptance into the archive, public dissemination. (From [11]).

Figure 7. Typical model rebuilding by PDBREDO. It should be emphasised that both Original and Optimised models are based on the *same* experimental data. (A) Changing the orientation of a peptide plane (*not* changing the peptide bond from trans to cis) achieves better fit to the electron density and gains a hydrogen bond. (B) Reconforming the distal region of a sidechain achieves better fit to the density. (C) Introducing atoms missing in the original structure makes clear that there are salt bridges. (D) Changing the conformation of the histidine and threonine sidehains does not achieve a significantly better fit to the electron density, but clarifies the hydrogen-bonding

pattern. From: [22].

Figure 8. The initial entry page for oxymyoglobin, 1mbo, in PDBe (`https://www.ebi.ac.uk/pdbe/entry/pdb/1mbo`). This page introduces the entry 1mbo, the X-ray structure of sperm whale oxymyoglobin refined at 1.6 Å resolution. (This is not the original structure published by by Kendrew and colleagues in 1958 (see [27]), but one of a subsequent generation of myoglobin structures, this one by S.E.V. Phillips [28].) The page summarises the following, and offers hypertext links to more details about this structure, and its relationships to others:

- General information: Structure, source, publication
- Function and Biology
- Structure analysis
- Ligands and Environments
- Experiments and Validation
- A 'cartoon' representation of the structure, subject to interactive reorientation

Figure 9. A few typical lines from the classic PDB format for the entry 1mbo, Sperm Whale Myoglobin, oxy from, solved by Simon E.V. Phillips. ... indicates omitted lines. Comments in brackets added for this article; they do not appear in the original file.
Note that each line is started by a keyword. An ATOM line specifies: keyword (= ATOM), residue, and chain identifier, x, y, and z coordinates, occupancy, and B-factor. For instance, atom CD1 of Tyr151 has two possible conformations, each with occupancy 0.5, and B-factors 21.83 and 27.93.
ATOM lines are fixed-format, in that data items appear in specific columns. That is, the format is *not* white-space independent. For Tyr151, some of the sidechain atoms have alternative conformations. The values of the B-factors – only slightly larger than those of atoms with single conformations only – suggest that the crystal contains two discrete well-defined conformations, that the sidechain is not merely 'flapping in the breeze'. (This could be checked using the electron-density server.)

Figure 10. The distinction between Architecture and Topology in CATH: two four-α-helix bundles with similar spatial arrangements of the helices but different connectivity. (a) A bundle with short, direct, connections between each successive pair of helices; that is, a succession of $\alpha-$hairpins, as in the GTPase domain of the signal recognition particle from *Thermus aquaticus,* from PDB entry 1ffh. Each successive pair of helices is antiparallel. The CATH CODE is 1.20.120.140; the 20 refers to the architecture 'Up-down bundle'. (b) A bundle with a 'bottom-to-top' connection (shown in red), as in the putative ATP-binding cobalamin adenosyltransferase YvqK, from PDB entry 1rty. The successive helices linked by the long connection are *parallel.* The CATH Code is 1.20.1200.10. These domains have the same Architecture but different Topology (after [68]).
Other proteins with similar spatial arrangements of secondary-structure elements but different connectivity appear in double-β-sheet sandwich proteins; for instance, immunoglobulin domains.

Figure 11. The Functional Families and functional annotations associated with HUP superfamily in CATH (v4.3). HUP superfamily is a highly diverse superfamily, comprising 922 FunFams in CATH v4.3.

(a) (a) CATH classification scheme for HUP superfamily. HUP superfamily is assigned CATH ID: 3.40.50.620.

(b) The superposition of representative structures (total 136) within the HUP superfamily, provided on the CATH webpage for this superfamily.

(c) Summary of Superfamily-level data and annotations

(d) Selection of some of the 922 Functional Families associated with HUP superfamily in CATH. For each FunFamCATH provides a Diversity Score (range 0–100) which reflects the information content in the multiple sequence alignment for the FunFam, as measured by scorecons [72].

Figure 12 Domain relationships in pyruvate decarboxylase and transketolase. Pyruvate decarboxylase converts pyruvate to acetaldehyde. Transketolase takes a ketose sugar and an aldose sugar, and converts the ketose to an aldose and the aldose to a ketose. Both enzymes use the cofactor thiamine pyrophosphate. Both (a) pyruvate decarboxylase and (b) transketolase contain three domains. They share two of the three domains, but the domains appear in different orders along the polypeptide chains. Nevertheless, the interface between the PYR and PP domains, containing the active site, is preserved between the two structures (c).

(a) Domain architecture of pyruvate decarboxylase, comprising PYR = pyrimidine ring binding domain (blue), TH3 = transhydrogenase dIII subunit (magenta), and PP = diphosphate binding domain (green). The cofactor thiamine pyrophosphate is shown in a shaded-sphere representation. (b) Domain architecture of transketolase, comprising PP (green), PYR (blue), and TKC = transketolase C-terminal domain (orange). (c) Pyruvate decarboxylase and transketolase, superposed on PYR and PP domains. Despite the difference in overall domain architecture in these two proteins, the geometric relationship between these two domains is preserved. The colours of the domains of pyruvate decarboxylase are the same as in the preceding figure: PYR blue, TH3 magenta, PP green. But in this figure the transketolase domains appear thus: PP red, PYR purple, TKC remains orange. The colours of the PP and PYR domains of transketolase have been changed in order to distinguish them from the superposed domains of pyruvate decarboxylase. (from [77], based on [78]).

Figure 13. SCOP family GH64 $\beta-$glucanase-like, of proteins containing two domains belonging to different superfamilies, different folds, and different classes. (From [74]. Reproduced by permission.)

Figure 14. (a) Prediction by AlphaFold2 of domain 2 of target T1038 from CASP 14, tomato spotted wilt tospovirus glycoprotein. (b) Prediction by AlphaFold2 of domain 1 of target T1049

from CASP 14, major virulence-associated fimbrial protein, MrpH, of the bacterial urinary tract pathogen *Proteus mirabilis.* In each case, cyan: experimental structure, magenta: prediction (from: [75].)