Published in final edited form as:

Methods Mol Biol. 2015 ; 1261: 3–20. doi:10.1007/978-1-4939-2230-7_1.

Protein Structure Annotation Resources

Margaret J. Gabanyi and

Center for Integrative Proteomics Research, Rutgers, The State University of New Jersey,
Piscataway, NJ 08854 USA.

Helen M. Berman

Center for Integrative Proteomics Research, Rutgers, The State University of New Jersey,
Piscataway, NJ 08854 USA.

Abstract

A key reason three-dimensional (3-D) protein structures are annotated with supporting or derived information is to understand the molecular basis of protein function. To this end, protein structure annotation databases curate key facts and observations, based on community-accepted standards, about the ~100,000 3-D experimental protein structures that enable further insight into the structure-function relationship. This review will introduce the primary structure repositories, databases, and value-added structural annotation databases, as well as the range of annotations they provide. The different levels of annotation data (primary vs. derived vs. inferred) will also be described and how they should all be considered accordingly.

Keywords

crystallography; databases; 3D electron microscopy; nuclear magnetic resonance; small angle scattering; structural biology; structural proteomics

1. Introduction

There is a vast amount of biological data available in public repositories. At the time this review was prepared, there were over 157 billion genes registered in GenBank (v. 200, Feb 2014) (1), half a million Swiss-Prot (curated) protein sequences and 54 million TrEMBL (computationally annotated) protein sequences in UniProtKB (v. 2014_03) (2), and nearly 100,000 three-dimensional macromolecular structures in the Protein Data Bank archive (v. 2014-03-25) (3, 4). Improved high-throughput technologies have helped gather new information faster than we are able to analyze it. There simply isn't enough time, funding, or skilled workforce to study every one. Therefore, to maximize the amount of information garnered from the genes, proteins, and structures we do study, annotations are captured in the hopes that they may apply to similar (*homologous*) proteins.

Since the 3-D structure of proteins are ultimately responsible for most actions in the cell, this review will focus on protein structure annotation resources that aim to explain a

molecular basis of protein function. First, we will describe the Protein Data Bank archive that houses the primary three-dimensional structural data and annotations for proteins and nucleic acids. We will then review the principal databases and resources that allow access to the PDB data for visualization and analysis, as well as several new repositories addressing new methods in structural biology. Next, structural classification and annotation databases that aim to categorize observations made by the structural biology community will be described. Selected specialty databases that focus on a specific biological aspect will also be covered. Lastly, we will review data aggregators and community-driven structure annotation projects that work to integrate protein structures with functional annotation to enable a better understanding of living systems and disease.

There is a hierarchy to the types of annotations that are captured and how reliable they are. At the top level is the structural data itself, composed usually of coordinates or 3-D maps. Next are the primary annotations, which have either been experimentally measured and verified, or report experimental parameters so that the study is reproducible. Next follow the derived annotations, which are calculated from experimental measurements. Last come the annotations that are propagated from homologous protein sequences or structures. It is important for biologists to understand the strengths and caveats of these data so that they can be applied to their maximum benefit. For this reason, it is important that the community experts be involved with the determination of these annotations to ensure consistency and reliability of the data for everyone.

2. The Protein Data Bank (PDB) Archive

The **PDB** is the primary worldwide data archive for experimentally-determined structures of biological macromolecules. It was established in 1971 following several years of discussion by the community who felt that such data should be openly available to the entire biological research community (5). At the 1971 Cold Spring Harbor Symposium *Structure and Function of Proteins at the Three-Dimensional Level*, Walter Hamilton of Brookhaven National Laboratory agreed to start the archive in collaboration with Olga Kennard of the Cambridge Structural Database; at that time there were less than a dozen protein structures. Now 40+ years later, there are almost 100,000 structures in the PDB and coordinates files are downloaded more than one million times every day.

Since the mid-2000s, the PDB archive has been managed by the worldwide Protein Data Bank (wwPDB) consortium (4), composed of four international members dedicated to maintaining a high-quality, uniformly curated resource: The Research Collaboratory for Structural Bioinformatics PDB (RCSB PDB, USA) (3), PDB in Europe (PDBe, UK) (6), PDB in Japan (PDBj, Japan) (7), and the BioMagResBank (BMRB, USA) (8). Being a community resource, the wwPDB regularly establishes task forces and meets with field experts who make recommendations to address growing complexities in the data. Recent developments made in response to these expert recommendations include the ongoing implementation of latest validation standards into data validation and processing procedures (9–11), the availability of structure validation reports for journal referees during peer-review (12), and, very recently, the shift towards the mmCIF/PDBx file format from the limiting PDB format (13).

The wwPDB collects structural data and also preliminary information about the macromolecules present in the structure (14) which is checked and represented using a data dictionary (15). For structures determined by X-ray crystallography, the wwPDB collects the x , y , z (Cartesian) coordinates for every resolved atom in a biomolecule, and the original structure factor data to generate the electron density maps. For structures determined by nuclear magnetic resonance (NMR), the Cartesian coordinates for each model in the deposited ensemble and restraint and chemical shift data are collected. In addition to the coordinates and data sets, methods-based annotations that represent how the sample was made and stored such as source and host organisms and vectors, crystallization conditions or NMR solutions are also included. Additional author-provided or calculated (derived) annotations include secondary structure element identification and symmetry and biological assembly information. Biological annotations such as GenBank and UniProt identifiers and the broad function of the macromolecule or enzyme classification are added to the record as well. Discrepancies between reported and reference sequences (such as mutations) are noted.

The next step in the annotation process is to validate the deposited atomic data. Covalent bond distances and angles, stereochemical validation, close contacts, ligand and atom nomenclature, sequence comparison, distant waters, and overall geometry are reviewed and compared by wwPDB annotators to standards recommended by the method's validation task force, and authors are informed of any inconsistencies. For X-ray structures, the fit of the refined model is also checked against its experimental data. Outliers may be noted in the entry. After data are fully reviewed and annotated they are approved by the depositing authors for public release. The PDB archive File Transfer Protocol (FTP) server is updated each week. The wwPDB regularly reviews the archive and makes corrections to data files to improve the quality and consistency of the data across the archive. These efforts also result in improved data processing procedures, and in some cases, the creation of external reference files and/or new data dictionaries. Recent work includes improving how peptide-like inhibitors and antibiotics are represented in the PDB archive (16). These are the first installment in a new "Biologically Interesting molecule Reference Dictionary" (BIRD), which contains chemical descriptions, sequence and linkage information, and functional and classification information as captured from the structures and other external sources.

It should be noted, however, that inclusion of other chemical components included in the entry do not infer any function. It is usually difficult to determine if the chemical components (such as those coming from a crystallization solution) are an artifact of the experiment or if they mimic some functional co-factor or substrate. The same is true for "site" records; authors can provide annotations regarding the functional role of an amino acid, but wwPDB also automatically calculates which amino acids are within a certain distance of nearby ligands and reports all of them.

A new unified Common Deposition and Annotation system has been developed by the wwPDB, and is being released to the community in 2014 (<http://deposit.wwpdb.org/>) (17). To address data quality issues, new validation metrics report how well each structure compares to those solved by the same method, and also how the structure compares to those determined to a similar resolution (9). Validation reports for all X-ray crystallographic

entries are now publicly available from the PDB FTP archive as of March 2014 (ftp://ftp.wwpdb.org/pub/pdb/validation_reports/).

3. Macromolecular Structure Resources

There are many data resources that depend wholly, or in part, on the contents of the PDB. This section will describe how each resource site utilizes the structural annotation information for search and analysis. The URL addresses for these resources are listed in Table 1.

The **RCSB PDB** provides structure search, visualization, and analysis tools to retrieve relevant entries from the PDB archive (18). Simple searches, including by PDB ID, molecule name, author name, or plain text are available. As words are entered in the search box, the autocomplete function will find individual structure entries, or suggest groups of structures that share a particular attribute or annotation (molecule type, structural fold, protein family, etc.). Users can also enter a sequence or draw a ligand structure, or use the ‘advanced’ and ‘browse’ search features to filter the archive by combining ~80 categories of structural (i.e., folds), biological (i.e. Gene Ontology (GO)), or experimental (i.e., method) annotations. Unless looking for a specific entry, the result is a list of relevant structures that can be organized into customizable reports for future reference.

An entry’s RCSB PDB Structure Summary page lists the previously mentioned annotations. Further primary and derived data such as secondary structure depiction, detailed methods information, sequence and structure similarity, protein domain and biological annotations, and many other data are arranged over 10 additional tabs to give a complete overview of the macromolecules observed in each structure entry. Additional residue level and domain level annotations are retrieved from UniProt and presented in a new “protein feature” view to show the structure-function relationships. Several molecular viewers (Jmol/JSmol, Simple Viewer, Protein Workshop) are integrated to visualize and maneuver the structure within the web browser. RCSB PDB’s Ligand Explorer also allows for detailed visualization of derived protein-ligand annotations such as involved side chains, and displays calculated bond lengths and interatomic distances (19). Several web-based structure analysis tools reduce the need for additional software. The Protein Comparison Tool gives access to six algorithms to calculate pairwise sequence or structure alignments (20). A Drug and Drug Mapping tool combines PDB structural data with **DrugBank** (21) annotations to locate all structures that are bound to pharmaceuticals or are known to act as drug targets.

RCSB PDB also hosts a large education and outreach corner called “PDB-101” that contains hands-on structure-related activities, educational posters, and structural biology tutorials. It also hosts over a decade’s worth of *Molecule of the Month* essays that teach a general audience how the shape of biologically important macromolecules and their annotated attributes (such as a residue’s role in a catalytic triad or pore selectivity filter) explains their function in living systems.

The **PDBe** provides access to the same PDB archive, but has developed different tools to find structures and group them by annotation. Upon entering a PDB ID of interest, it gives

“one-click” access to the structural entry and their annotations in PDBeAtlas, the ability to download the files, predict protein interfaces and quaternary assembly of the protein using PDBePISA (22), find similarly-folded structures using PDBeFold/SSM (23), or find certain structural motifs or binding sites within each structural chain via the tool PDBeMotif (24). They have also developed a new structure browser based on GO annotations (6). The PDBe is an active contributor to the SIFTS annotation and mapping project which works to integrate taxonomy, protein sequences, structure, and function (25).

The **PDBj** provides similar structure search, visualization, and analysis tools, and it is also the only member site that supports browsing in additional languages such as Japanese, simplified/traditional Chinese, and Korean. The SeSAW tool identifies functionally or evolutionarily conserved motifs in protein structures by locating sequence and structural similarities, and annotating these at the residue level (26). It has also built tools for bioinformaticians and, in collaboration with the RCSB PDB group, has developed an XML version of the PDB archive (PDBML) (7) and an extended version (PDBMLplus) with additional curated annotations that can be searched using the PDBj Mine application (27).

The **BMRB** became a member of the wwPDB in 2006 (28). It is a repository that collects NMR data from any experiment, not only those focused on determining structures. It captures the assigned chemical shifts, coupling constants, and peak lists for a variety of biological macromolecules (even small ones excluded from the PDB). It also contains derived annotations such as hydrogen exchange rates, pKa values, and relaxation parameters that explain biochemical processes in real-time. BMRB also collects the NMR restraints for PDB entries, time domain spectral data, and NMR data on hundreds of metabolites and standard compounds (8).

Larger biological assemblies are increasingly studied via three-dimensional electron microscopy (3DEM) methods; these data are available from the **EMDatabank** resource, a joint project of the National Center for Macromolecular Imaging (NCMI), PDBe and the RCSB (29). Both maps and model coordinates fitted into the maps are accessible via EMDatabank services. EMDatabank annotations include sample description, sample components, stoichiometry, specimen preparation, details about the imaging experiment, image processing, reconstruction method, resolution, fitting description and source PDB IDs used in fitting the map. Since 3DEM structures are typically large and complex biological assemblies such as viruses and ribosomes, EMDatabank's sample description annotation permits hierarchical description of the complex components, with the capability to drill down to entity level. Search capabilities are provided at the EMDatabank site.

The **Nucleic Acid Database** (NDB) provides access to information about three-dimensional nucleic acid structures and their complexes (30). In addition to principal coordinate data, the NDB contains nucleic acid-specific annotation and derived structural and geometric data. Nucleic acid structures are deposited to the wwPDB since they are biological macromolecules, and as such, many of the same annotations such as coordinate data, experimental parameters, identifiers, structural description, and data collection and refinement details are captured. Nucleic acid specific annotations such as conformation type, secondary structure information, nucleic acid type, and also protein/drug binding partner

information are curated and stored in an ERF that is searchable from the NDB website. Due to recent growth of RNA biology, the NDB was recently redesigned to include additional derived information on RNA structural features such as pairwise nucleotide interactions, base-stacking interactions, base-phosphate interactions, sequence and geometry-based equivalence classes containing similar structures, new clustering of non-redundant RNA structure sets, and an “atlas” of representative RNA 3D motifs as extracted from the structures. Users can then search the NDB structures using the aforementioned annotations as constraints, create reports, and download them for further analysis.

Another source of protein structure information is the emerging field of Small Angle Scattering (SAS). These methods are growing in popularity because most proteins are not solely comprised of globular domains; many functional sites are known to be on transiently structured or intrinsically-disordered portions of the protein that cannot be visualized by x-ray crystallography or are too large for NMR studies. There are two repositories that currently collect such data: **BioIsis** (SAXS) (31) and the **Protein Ensembles Database** (pe-DB) (NMR and SAS) (32). Currently their annotations are limited to capturing experimental parameters and mapping identifiers to UniProt for functional information, but active communities are in the process of further developing those annotation dictionaries (33).

4. Other Structural Annotation Resources

In the early days of protein structure determination, researchers noted that polypeptides would assemble secondary structure elements into regular 3-D patterns (folds) that could be categorized. After many more protein structures were determined, new structures were composed of different combinations of these folds rather than new folds. Those interested more in the biological aspects created resources focused on specific protein superfamilies. This section first describes the databases that classify and annotate the tertiary arrangement of the proteins. It will also briefly summarize the resources that leverage this tertiary information to align similar structures and cluster them. The URL addresses are listed in Table 1.

The Structural Classification of Proteins (**SCOP**) database classifies structures in the PDB in a hierarchical fashion: ‘family’, ‘superfamily’, ‘common fold’ and ‘class’ (34). A ‘*family*’ is determined by sequence similarity. A ‘*superfamily*’ is a set of families with similar structure and function. Families and superfamilies with the same arrangement of secondary structures and connectivity have the same ‘*common fold*’. The types of secondary structures in these folds are ‘*classes*’ (all alpha helix, all beta sheet, alpha-beta, etc.). This type of analysis connects structures with possible evolutionary relationships. One limitation of SCOP is that the current version (1.75) was released in 2009 and contained ~one-third of today’s PDB archive. The SCOP group first addressed this gap by releasing SCOP-extended (**SCOPE**) that classifies structures added to the archive since 2009 in an automated fashion (35). It currently contains 176,192 domains, 4,496 families, 1,961 superfamilies, and 1,194 folds from 62,616 PDB entries (as of the Feb 2014 release). Further review of the data showed that the original hierarchy was too simple, and a new database SCOP2 was created and will be released in 2014 (36). The proteins are still annotated according to their structural and thus inferred evolutionary relationships, but due to the increase in structure data, the tree-like

hierarchy had to be expanded to a complex network of nodes. The relationships are now organized into four new categories: protein type, evolutionary events, structural classes, and protein relationships, with mostly new content and definitions and some carryover content from SCOP. The SCOP2 prototype currently encompasses only a select portion of the PDB archive, but this is expected to expand once it is released.

The **CATH** database uses a similar classification scheme based on class (C), architecture (A), topology (T) and homologous superfamilies (H) (37). *Class* describes the secondary structure content as in SCOP. *Architecture* defines the description of the arrangement of these secondary structures without consideration of the connectivities. *Topology* is equivalent to fold in SCOP. Finally, *homologous superfamilies* contain all folds with a similar function. In its current release (v3.5, version from September 2011), it has annotated 173,500 CATH domains, over 2,600 CATH Superfamilies, and 1,300+ fold groups from an analysis of the ~51,000 PDB entries that were in the archive in 2011. Functional sub-classifications (FunFams) have been added to the homologous superfamily layer so that functional divergence within a superfamily could be better understood (38).

Structural annotations such as those curated by CATH and SCOP are being leveraged to help close the sequence-structure gap. **SUPERFAMILY**, part of the SCOP family of databases, organizes structures into evolutionarily-related groups to promote the annotation of under-characterized protein (39). **Gene3D**, part of the CATH family of databases, plays a similar role in assigning CATH domains to gene products of unknown structure (40). The **Genome3D** project is a collaborative project between SCOP, CATH, and the leading structure prediction resources to extend structural domain and 3D structure information to more model genomes, and in the process serve as the first official mapping between SCOP and CATH (41).

The globular fold is not the only functional interface for proteins, as many binding sites appear on variable loops between α -helices and β -strands. The **ArchDB** classification database classifies loops extracted from structures in the PDB archive (42). These loops are annotated based on their length, its ϕ and ψ backbone dihedral angles, the type of flanking secondary structures, the distance between the loop ends, and how the loop is supported (braced) by nearby secondary structure elements. The current version of ArchDB (v 2.0.0) contains annotations for 149,134 loops in 5739 classes and 9608 subclasses. These data are helpful for the theoretical modeling and protein design communities.

Sometimes the structural classification is that there is –no- shape, with portions of the protein adopting a random coil configuration. The new **MobiDB** database contains the most comprehensive data on intrinsically-disordered proteins (43). MobiDB contains the 600+ manually curated entries from **DisProt** (44), extends the data with additional evidence of unresolved residues from the PDB (both X-ray and NMR data), adds disordered region predictions from nine bioinformatics resources, and overlays biologically relevant annotations within the regions regarding tissue localization, function, and residue-level annotations from UniProt.

The reason macromolecular structures are annotated with supporting or derived information is to understand the molecular basis of protein function. As a result there are as many specialty structure annotation databases as there are protein families or structure-function attributes. It is beyond the scope of this paper to cover all of them, so we direct the reader to the Nucleic Acids Research online catalog of protein structure resources at <http://www.oxfordjournals.org/nar/database/subcat/4/14>. This second portion therefore, will focus only on two major areas of current research: protein-protein interaction databases and membrane protein annotation databases. The URL addresses for the resources described herein are listed in Table 2.

A protein is rarely isolated in the cell, and sometimes self-associates or binds to other subunits to become functional. This is not always evident in the structure entry. While not strictly annotation databases, there are many tools that derive these essential 3-D data (and thus deriving new annotations). **PDBePISA** can determine the quaternary assembly of macromolecules derived from calculations made upon the molecule's surface or observed interfaces (45). The **Dictionary of Interfaces in Proteins** (DIP) is a data bank of complementary molecular surface patches and is meant to enable molecular recognition research (46). Other bioinformatics tools like **ProtBuD** (47) analyze the surfaces of proteins in the PDB archive to review interactions seen in the crystallographic lattice and to equate them to the likelihood of being the correct biological assembly. A new database of three-dimensional interacting domains (**3did**) annotates binding modes observed in domain-domain interactions and also in domain-peptide interactions from the PDB archive (48). It has identified 8,651 interactions and classified them into 483 motifs, and is updated twice a year.

Another area of intense research is membrane proteins, since they play key roles in controlling the processes of life. Both the Membrane Proteins of Known Structures database (**mpstruc**) (49) and the Protein Data Bank of Transmembrane Proteins (**PDBTM**) (50) annotate entries with tertiary structure information along with their orientation, topology, and assembly in the membrane. Historically, mpstruc was manually curated by biological superfamily through reports from the literature, but a recent collaboration with **OMPdb** (51) and the RCSB extended the list of membrane proteins to nearly ~1700 entries. The GPCR database (**GPCRdb**) focuses only on the G-protein coupled receptor superfamily, an important drug target. It integrates residue-level mutation and ligand binding data with their sequences, and structures when available (52). The Transporter Classification Database (**TCDB**) uses a five-level classification scheme to organize phylogenetic and functional information together (53). The classification includes transporter class, subclass (such as energy source to drive transport), transporter family/superfamily, subfamily, and the range of substrates transported, and also includes filters for human transporters and transporters connected to disease states. If an experimental structure is not available, the TCDB link to external tools that can look for orthologous protein structures or predict transmembrane segments.

5. Data Aggregators

The volume of structural and biological data continues to expand rapidly, and it typically spans over several orders of resolution from molecular mechanisms to cellular phenotypes and tissues. This makes it difficult for non-bioinformaticians to find remote connections between structure and function. Large bioinformatics programs like NCBI (Entrez) and EBI (EBI Search) have search portals that will search all of their underlying databases and return a inventory list of related but disjointed entries. This section describes a structure-centric data aggregator, the **Structural Biology Knowledgebase (SBKB)**, which accesses hundreds of annotation databases under the hood. The URL addresses for resources described in this section are found in Table 2.

The SBKB was established as a scientific portal to facilitate research design and analysis for a wide variety of biological systems (54). It serves as a single resource that integrates structure, sequence, and functional annotations as well as technical information regarding protein production and structure determination. Researchers can search the SBKB by sequence, PDB ID or UniProt accession code, and receive an up-to-date list of matching 3D experimental structures from the PDB, pre-built theoretical models federated by the **Protein Model Portal** (55), annotations from 100+ open genomic, protein, structural, and functional resources, Protein Structure Initiative (PSI) structural genomics target histories and protocols from **TargetTrack** (56), and ready-to-use DNA clones from **DNASU** (57). It is also possible to find structures by text according to functional and disease relevance (**KB-Rank tool**) (58), or find related technologies and publications from the **PSI Technology Portal** (59) and **PSI Publications Portal**, respectively. Interactive tools such as real-time theoretical modeling and biophysical parameter prediction also enhance understanding of proteins that are not yet well characterized. A Functional Sleuth viewer collects a list of under characterized proteins that require further study, updated on a weekly basis.

To act as a research concierge, the SBKB has created five “**Hubs**” that collect practical information about Structural Targets, Modeling, Membrane Proteins, Methods, and ‘Structure, Sequence, and Function’ from its portal modules and the World Wide Web. Select examples include the Membrane Protein Structure Hub which has links to methods for membrane protein production, the latest results from the Protein Structure Initiative’s Human Transmembrane Proteome Coverage project (60), and links to several public membrane protein structure resources. The ‘Structure, Sequence, and Function’ Hub lists all biological resources used during an SBKB query, and provides a summary of the search and analysis features for each resource. We refer the reader to this Hub for information on additional resources that could not be included in this review.

6. Community-driven Structure Annotation Projects

Due to the success of open annotation platforms such Wikipedia, some groups are taking the same collaborative approach to gather expert information and discussion on scientific topics, including protein structures. This final section will describe recent community annotation efforts for protein structure annotation. The URL addresses for the resources mentioned in this section are found in Table 2.

Proteopedia is web-based structure “wiki” that visualizes molecules and their highlighted structural or functional properties (61). It allows any registered user to edit any page, and added hyperlinks in the narrative text can trigger the molecular viewer to rotate to show the highlighted structural feature at its best angle. Every PDB entry has a page, which at first contains automatically retrieved text from multiple sources such as **ConSurf**, which identifies functional regions in proteins (62), and **OCA**, a structure-function browser (63). Higher-tier articles called “topic pages” describe the protein or protein families and link to individual entry pages. Users can also create the view and export it into Microsoft PowerPoint for teaching purposes. Author-contributed and peer-reviewed Proteopedia pages are also indexed in PubMed.

The Open Protein Structure Annotation Network (**TOPSAN**) is a collaborative annotation platform focused on collecting annotations on structures solved by high-throughput structural genomics efforts (64). In many cases, their proteins were selected and structurally determined before any functional information was available, so annotating the structure with value-added information presents a challenge. Similar to Proteopedia, a registered contributor can edit a page, and have it include additional author-created view of the structure and supplemental data in the form of file attachments. The same group at the Joint Center for Structural Genomics has hosted Domain of Unknown Function (DUF) “jamborees” where top bioinformatics experts collaborate to annotate protein families that are listed as having unknown functions; the effort resulted in the discovery of new Pfam domains, some of which have since been included in Pfam.

7. Getting the Most Value Out of Value-Added Annotations

All structural resources have been set up with a common goal – to understand the complex relationship between protein sequence, structure, and function. As indicated in this review, some annotations are experimentally determined, while other value-added annotations are derived or inferred by homology. Rather than keeping the data in silos, all biological (including structural) databases are beginning to integrate cross-referenced annotations from the other resources to make the sequence, structure, and function connections more evident. However, the origins of the annotations (primary vs. derived) are sometimes buried in documentation (or worse, ignored) and then annotations that were originally inferred by homology are suddenly understood to be fact. There can also be a lag in time between the release of a new study and its incorporation, or clarification of previous annotations, into the databases. This leads to a cyclical propagation of errors, which can linger when some databases do not release updates as often as UniProt (~monthly) or PDB (weekly). There are plenty of articles that point out errors in the reference databases (65–67), including a recent case study of the repair of a mis-annotation by the UniProtKB Consortium (68). To the credit of the annotation databases, there is an immense amount of sequence data that needs to be covered and they are keen to quickly repair inconsistencies. But as data-users, we must be aware of the caveats that are implied in derived data, and to always check (and cite) the original source, so that validated science is always propagated.

Acknowledgments

The authors are grateful to B. Coimbatore Narayanan, G. Gabanyi, C. Lawson, E. Peisach, M. Quesada, M. Sekharan, J. Westbrook, and C. Zardecki for helpful discussions and review during the preparation of this work. The RCSB PDB (H.M.B.) is supported by the National Science Foundation (NSF DBI 0829586), the Department of Energy, the National Library of Medicine, the National Cancer Institute, National Institute of Neurological Disorders and Stroke, and the National Institute of Diabetes and Digestive and Kidney Diseases. The SBKB (H.M.B./M.G.) is supported by a grant from the National Institute of General Medical Sciences of the National Institutes of Health (U01 GM093324).

References

1. Benson DA, Clark K, Karsch-Mizrachi I, et al. GenBank. *Nucleic Acids Res.* 2014; 42:D32–37. DOI: 10.1093/nar/gkt1030 [PubMed: 24217914]
2. The UniProt Consortium. Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* 2014; 42:D191–198. DOI: 10.1093/nar/gkt1140 [PubMed: 24253303]
3. Berman HM, Westbrook JD, Feng Z, et al. The Protein Data Bank. *Nucleic Acids Res.* 2000; 28:235–242. [PubMed: 10592235]
4. Berman HM, Henrick K, Nakamura H. Announcing the worldwide Protein Data Bank. *Nat Struct Biol.* 2003; 10:980.doi: 10.1038/nsb1203-980 [PubMed: 14634627]
5. The Protein Data Bank. Protein Data Bank. *Nature New Biol.* 1971; 233:223.doi: 10.1038/newbio233223b0 [PubMed: 20480989]
6. Gutmanas A, Alhroub Y, Battle GM, et al. PDBe: Protein Data Bank in Europe. *Nucleic Acids Res.* 2014; 42:D285–291. DOI: 10.1093/nar/gkt1180 [PubMed: 24288376]
7. Kinjo AR, Suzuki H, Yamashita R, et al. Protein Data Bank Japan (PDBj): maintaining a structural data archive and resource description framework format. *Nucleic Acids Res.* 2012; 40:D453–460. DOI: 10.1093/nar/gkr811 [PubMed: 21976737]
8. Ulrich EL, Akutsu H, Doreleijers JF, et al. BioMagResBank. *Nucleic Acids Res.* 2008; 36:D402–408. DOI: 10.1093/nar/gkm957 [PubMed: 17984079]
9. Read RJ, Adams PD, Arendall WB III, et al. A new generation of crystallographic validation tools for the Protein Data Bank. *Structure.* 2011; 19:1395–1412. DOI: 10.1016/j.str.2011.08.006 [PubMed: 22000512]
10. Montelione GT, Nilges M, Bax A, et al. Recommendations of the wwPDB NMR Validation Task Force. *Structure.* 2013; 21:1563–1570. DOI: 10.1016/j.str.2013.07.021 [PubMed: 24010715]
11. Henderson R, Sali A, Baker ML, et al. Outcome of the first electron microscopy validation task force meeting. *Structure.* 2012; 20:205–214. DOI: 10.1016/j.str.2011.12.014 [PubMed: 22325770]
12. New wwPDB X-ray structure validation reports support depositors, journal editors and referees. The wwPDB Consortium; 2013. http://wwpdb.org/news/news_2013.html#02-August-2013 [Accessed March 1 2014]
13. Deposition and Release of PDB Entries Containing Large Structures. The wwPDB Consortium; 2013. http://wwpdb.org/news/news_2013.html#22-May-2013 [Accessed March 1 2014]
14. wwPDB Processing Procedures and Policies Document: Section B: wwPDB Policies. The wwPDB Annotation Staff; 2014. <http://www.wwpdb.org/policy.html> [Accessed March 1 2014]
15. Westbrook, JD., Fitzgerald, PMD. Chapter 10 The PDB format, mmCIF formats, and other data formats. In: Bourne, PE., Gu, J., editors. *Structural Bioinformatics*. Second. John Wiley & Sons, Inc; Hoboken, NJ: 2009. p. 271-291.
16. Dutta S, Dimitropoulos D, Feng Z, et al. Improving the representation of peptide-like inhibitor and antibiotic molecules in the Protein Data Bank. *Biopolymers.* 2014; 101:659–668. DOI: 10.1002/bip.22434 [PubMed: 24173824]
17. Quesada M, Westbrook J, Oldfield T, et al. The wwPDB common tool for deposition and annotation. *Acta Cryst.* 2011; C403–C404.
18. Rose PW, Bi C, Bluhm WF, et al. The RCSB Protein Data Bank: new resources for research and education. *Nucleic Acids Res.* 2013; 41:D475–D482. DOI: 10.1093/nar/gks1200 [PubMed: 23193259]

19. Moreland JL, Gramada A, Buzko OV, et al. The Molecular Biology Toolkit (MBT): a modular platform for developing molecular visualization applications. *BMC Bioinformatics*. 2005; 6:21.doi: 10.1186/1471-2105-6-21 [PubMed: 15694009]
20. Prlic A, Bliven S, Rose PW, et al. Pre-calculated protein structure alignments at the RCSB PDB website. *Bioinformatics*. 2010; 26:2983–2985. DOI: 10.1093/bioinformatics/btq572 [PubMed: 20937596]
21. Knox C, Law V, Jewison T, et al. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res*. 2011; 39:D1035–1041. DOI: 10.1093/nar/gkq1126 [PubMed: 21059682]
22. Krissinel, E., Henrick, K. Detection of Protein Assemblies in Crystals. In: Berthold, MR.Glen, R.Diederichs, K.Kohlbacher, O., Fischer, I., editors. *Computational Life Sciences; First International Symposium, CompLife 2005; Konstanz, Germany. September 25-27, 2005; Berlin: Springer-Verlag; 2005. p. 163-174.Proceedings*
23. Krissinel E, Henrick K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr D Biol Crystallogr*. 2004; 60:2256–2268. DOI: 10.1107/S0907444904026460 [PubMed: 15572779]
24. Golovin A, Henrick K. Chemical substructure search in SQL. *J Chem Inf Model*. 2009; 49:22–27. DOI: 10.1021/ci8003013 [PubMed: 19072559]
25. Velankar S, Dana JM, Jacobsen J, et al. SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. *Nucleic Acids Res*. 2013; 41:D483–489. DOI: 10.1093/nar/gks1258 [PubMed: 23203869]
26. Standley DM, Yamashita R, Kinjo AR, et al. SeSAW: balancing sequence and structural information in protein functional mapping. *Bioinformatics*. 2010; 26:1258–1259. DOI: 10.1093/bioinformatics/btq116 [PubMed: 20299324]
27. Kinjo AR, Yamashita R, Nakamura H. PDBj Mine: design and implementation of relational database interface for Protein Data Bank Japan. *Database (Oxford)*. 2010; 2010:baq021.doi: 10.1093/database/baq021 [PubMed: 20798081]
28. Markley JL, Ulrich EL, Berman HM, et al. BioMagResBank (BMRB) as a partner in the Worldwide Protein Data Bank (wwPDB): new policies affecting biomolecular NMR depositions. *Journal of Biomolecular NMR*. 2008; 40:153–155. [PubMed: 18288446]
29. Lawson CL, Baker ML, Best C, et al. EMDatabank.org: unified data resource for CryoEM. *Nucleic acids research*. 2011; 39:D456–464. DOI: 10.1093/nar/gkq880 [PubMed: 20935055]
30. Coimbatore Narayanan B, Westbrook J, Ghosh S, et al. The Nucleic Acid Database: new features and capabilities. *Nucleic Acids Res*. 2014; 42:D114–122. DOI: 10.1093/nar/gkt980 [PubMed: 24185695]
31. Hura GL, Menon AL, Hammel M, et al. Robust, high-throughput solution structural analyses by small angle X-ray scattering (SAXS). *Nat Methods*. 2009; 6:606–612. DOI: 10.1038/nmeth.1353 [PubMed: 19620974]
32. Varadi M, Kosol S, Lebrun P, et al. pE-DB: a database of structural ensembles of intrinsically disordered and of unfolded proteins. *Nucleic Acids Res*. 2014; 42:D326–335. DOI: 10.1093/nar/gkt960 [PubMed: 24174539]
33. Trewhella J, Hendrickson WA, Sato M, et al. Meeting Report of the wwPDB Small-Angle Scattering Task Force: Data Requirements for Biomolecular Modeling and the PDB Structure. 2013; 21:875–881.
34. Andreeva A, Howorth D, Brenner SE, et al. SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res*. 2004; 32:D226–229. [PubMed: 14681400]
35. Fox NK, Brenner SE, Chandonia JM. SCOPe: Structural Classification of Proteins--extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res*. 2014; 42:D304–309. DOI: 10.1093/nar/gkt1240 [PubMed: 24304899]
36. Andreeva A, Howorth D, Chothia C, et al. SCOP2 prototype: a new approach to protein structure mining. *Nucleic Acids Res*. 2014; 42:D310–314. DOI: 10.1093/nar/gkt1242 [PubMed: 24293656]
37. Cuff AL, Sillitoe I, Lewis T, et al. The CATH classification revisited--architectures reviewed and new ways to characterize structural divergence in superfamilies. *Nucleic Acids Res*. 2009; 37:D310–314. DOI: 10.1093/nar/gkn877 [PubMed: 18996897]

38. Sillitoe I, Cuff AL, Dessailly BH, et al. New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. *Nucleic Acids Res.* 2013; 41:D490–498. DOI: 10.1093/nar/gks1211 [PubMed: 23203873]
39. Wilson D, Madera M, Vogel C, et al. The SUPERFAMILY database in 2007: families and functions. *Nucleic Acids Res.* 2007; 35:D308–313. [PubMed: 17098927]
40. Lees JG, Lee D, Studer RA, et al. Gene3D: Multi-domain annotations for protein sequence and comparative genome analysis. *Nucleic Acids Res.* 2014; 42:D240–245. DOI: 10.1093/nar/gkt1205 [PubMed: 24270792]
41. Lewis TE, Sillitoe I, Andreeva A, et al. Genome3D: a UK collaborative project to annotate genomic sequences with predicted 3D structures based on SCOP and CATH domains. *Nucleic Acids Res.* 2013; 41:D499–507. DOI: 10.1093/nar/gks1266 [PubMed: 23203986]
42. Bonet J, Planas-Iglesias J, Garcia-Garcia J, et al. ArchDB 2014: structural classification of loops in proteins. *Nucleic Acids Res.* 2014; 42:D315–319. DOI: 10.1093/nar/gkt1189 [PubMed: 24265221]
43. Di Domenico T, Walsh I, Martin AJ, et al. MobiDB: a comprehensive database of intrinsic protein disorder annotations. *Bioinformatics.* 2012; 28:2080–2081. DOI: 10.1093/bioinformatics/bts327 [PubMed: 22661649]
44. Sickmeier M, Hamilton JA, LeGall T, et al. DisProt: the Database of Disordered Proteins. *Nucleic Acids Res.* 2007; 35:D786–793. DOI: 10.1093/nar/gkl893 [PubMed: 17145717]
45. Krissinel E, Henrick K. Inference of macromolecular assemblies from crystalline state. *J Mol Biol.* 2007; 372:774–797. DOI: 10.1016/j.jmb.2007.05.022 [PubMed: 17681537]
46. Salwinski L, Miller CS, Smith AJ, et al. The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.* 2004; 32:D449–451. [doi]. DOI: 10.1093/nar/gkh086 [PubMed: 14681454]
47. Xu Q, Canutescu A, Obradovic Z, et al. ProtBuD: a database of biological unit structures of protein families and superfamilies. *Bioinformatics.* 2006; 22:2876–2882. DOI: 10.1093/bioinformatics/btl490 [PubMed: 17018535]
48. Mosca R, Ceol A, Stein A, et al. 3did: a catalog of domain-based interactions of known three-dimensional structure. *Nucleic Acids Res.* 2014; 42:D374–379. DOI: 10.1093/nar/gkt887 [PubMed: 24081580]
49. Snider C, Jayasinghe S, Hristova K, et al. MPEx: a tool for exploring membrane proteins. *Protein Sci.* 2009; 18:2624–2628. DOI: 10.1002/pro.256 [PubMed: 19785006]
50. Kozma D, Simon I, Tusnady GE. PDBTM: Protein Data Bank of transmembrane proteins after 8 years. *Nucleic Acids Res.* 2013; 41:D524–529. DOI: 10.1093/nar/gks1169 [PubMed: 23203988]
51. Tsirigos KD, Bagos PG, Hamodrakas SJ. OMPdb: a database of {beta}-barrel outer membrane proteins from Gram-negative bacteria. *Nucleic Acids Res.* 2011; 39:D324–331. DOI: 10.1093/nar/gkq863 [PubMed: 20952406]
52. Isberg V, Vroiling B, van der Kant R, et al. GPCRDB: an information system for G protein-coupled receptors. *Nucleic Acids Res.* 2014; 42:D422–425. DOI: 10.1093/nar/gkt1255 [PubMed: 24304901]
53. Saier MH Jr, Reddy VS, Tamang DG, et al. The transporter classification database. *Nucleic Acids Res.* 2014; 42:D251–258. DOI: 10.1093/nar/gkt1097 [PubMed: 24225317]
54. Gabanyi MJ, Adams PD, Arnold K, et al. The Structural Biology Knowledgebase: a portal to protein structures, sequences, functions, and methods. *J Struct Funct Genomics.* 2011; 12:45–54. DOI: 10.1007/s10969-011-9106-2 [PubMed: 21472436]
55. Haas J, Roth S, Arnold K, et al. The Protein Model Portal--a comprehensive resource for protein structure and model information. *Database.* 2013 Apr 26.2013:bat031.doi: 10.1093/database/bat031 [PubMed: 23624946]
56. Chen L, Oughtred R, Berman HM, et al. TargetDB: A target registration database for structural genomics projects. *Bioinformatics.* 2004; 20:2860–2862. DOI: 10.1093/bioinformatics/bth300 [PubMed: 15130928]
57. Seiler CY, Park JG, Sharma A, et al. DNASU plasmid and PSI: Biology-Materials repositories: resources to accelerate biological research. *Nucleic Acids Res.* 2014; 42:D1253–1260. DOI: 10.1093/nar/gkt1060 [PubMed: 24225319]

58. Julfayev ES, McLaughlin RJ, Tao YP, et al. KB-Rank: efficient protein structure and functional annotation identification via text query. *Journal of structural and functional genomics*. 2012; 13:101–110. DOI: 10.1007/s10969-012-9125-7 [PubMed: 22270457]
59. Gifford LK, Carter LG, Gabanyi MJ, et al. The Protein Structure Initiative Structural Biology Knowledgebase Technology Portal: a structural biology web resource. *Journal of structural and functional genomics*. 2012; 13:57–62. DOI: 10.1007/s10969-012-9133-7 [PubMed: 22527514]
60. Pieper U, Schlessinger A, Kloppmann E, et al. Coordinating the impact of structural genomics on the human alpha-helical transmembrane proteome. *Nat Struct Mol Biol*. 2013; 20:135–138. DOI: 10.1038/nsmb.2508 [PubMed: 23381628]
61. Prilusky J, Hodis E, Canner D, et al. Proteopedia: a status report on the collaborative, 3D web-encyclopedia of proteins and other biomolecules. *J Struct Biol*. 2011; 175:244–252. DOI: 10.1016/j.jsb.2011.04.011 [PubMed: 21536137]
62. Ashkenazy H, Erez E, Martz E, et al. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res*. 2010; 38:W529–533. DOI: 10.1093/nar/gkq399 [PubMed: 20478830]
63. Prilusky, J. [Accessed March 1 2014] OCA, a browser-database for protein structure/function. 1996. <http://oca.weizmann.ac.il>
64. Krishna SS, Weekes D, Bakolitsa C, et al. TOPSAN: use of a collaborative environment for annotating, analyzing and disseminating data on JCSG and PSI structures. *Acta Crystallogr Sect F Struct Biol Cryst Commun*. 2010; 66:1143–1147.
65. Zheng H, Chordia MD, Cooper DR, et al. Validation of metal-binding sites in macromolecular structures with the CheckMyMetal web server. *Nat Protoc*. 2014; 9:156–170. DOI: 10.1038/nprot.2013.172 [PubMed: 24356774]
66. Richardson CR, Luo QJ, Gontcharova V, et al. Analysis of antisense expression by whole genome tiling microarrays and siRNAs suggests mis-annotation of Arabidopsis orphan protein-coding genes. *PLoS One*. 2010; 5:e10710.doi: 10.1371/journal.pone.0010710 [PubMed: 20520764]
67. Schnoes AM, Brown SD, Dodevski I, et al. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput Biol*. 2009; 5:e1000605.doi: 10.1371/journal.pcbi.1000605 [PubMed: 20011109]
68. Poux S, Magrane M, Arighi CN, et al. Expert curation in UniProtKB: a case study on dealing with conflicting and erroneous data. *Database (Oxford)*. 2014; 2014:bau016.doi: 10.1093/database/bau016 [PubMed: 24622611]

Table 1

List of structural resource selected for this review. URL web addresses are provided and brief summaries describe the scope of each database.

Repository for 3D Biological Macromolecule Structures		
Protein Data Bank (3)	macromolecule structure coordinates archive	http://www.wwpdb.org
Structure Resources		
RCSB Protein Data Bank (3)	macromolecular structure deposition, search, and analysis tools (USA)	http://www.rcsb.org
PDB in Europe (PDBe) (6)	macromolecular structure deposition, search, and analysis tools (Europe)	http://www.pdbe.org
PDB in Japan (PDBj) (7)	macromolecular structure deposition, search, and analysis tools (Japan)	http://www.pdbj.org
BioMagResBank (BMRB) (8)	NMR data set deposition, search, and validation tools	http://www.bmrb.wisc.edu
EMDataBank (29)	Cryo-EM data deposition, search, and analysis tools	http://www.emdatabank.org
Nucleic Acid Database (NDB) (30)	nucleic acid structure search and analysis tools	http://ndbserver.rutgers.edu
BioIsis (31)	SAS data deposition	http://www.bioisis.net/
Protein Ensemble Database (pe-DB) (32)	NMR and SAXS data deposition for intrinsically-disordered proteins	http://pedb.vib.be
Structural Annotation Resources		
SCOP suite (SCOPE, SCOP2) (34–36)	structural classification	http://scop.mrc-lmb.cam.ac.uk/scop/
CATH (37)	structural classification	http://www.cathdb.info
SUPERFAMILY (39)	structural alignment and annotation	http://supfam.cs.bris.ac.uk/SUPERFAMILY/
Gene3D (40)	structural alignment and extended annotation	http://gene3d.biochem.ucl.ac.uk/Gene3D/
Genome3D (41)	structural alignment and extended annotation	http://genome3d.eu/
ArchDB (42)	structural classification of loops	http://sbi.imim.es/archdb/
MobiDB (43)	intrinsically-disordered proteins database	http://mobidb.bio.unipd.it
DisProt (44)	intrinsically-disordered proteins database	http://www.disprot.org

Table 2

Continued list of other annotation databases, data aggregators, and community-driven annotation resources selected for this review. URL web addresses are provided and brief summaries describe the scope of each database.

Other Structural Annotations Databases		
PDBePISA (45)	intermolecular interactions and quaternary structure	http://pdbe.org/pisa
Dictionary of Interfaces in Proteins (DIP) (46)	protein-protein interactions	http://dip.doe-mbi.ucla.edu/dip/Main.cgi
ProtBuD (47)	known and predicting protein interactions	http://dunbrack.fccc.edu/ProtBuD/
3did (48)	domain-domain and domain-peptide interactions	http://3did.irbbarcelona.org/
Membrane Proteins of Known Structure (mpstruc) (49)	membrane protein structure search and annotation	http://blanco.biomol.uci.edu/mpstruc/listAll/list
Protein Data Bank of Transmembrane Proteins (PDBTM) (50)	membrane protein structure search and annotation	http://pdbtm.enzim.hu/
GPCRDB (52)	G-protein coupled receptor protein structure search and annotation	http://www.gpcr.org/7tm/
Transporter Classification Database (53)	classification of membrane transport proteins	http://www.tcdb.org
Data Aggregators		
Structural Biology Knowledgebase (SBKB) (54)	structure-centric data aggregator and search tool for sites below	http://sbkb.org
PSI Protein Model Portal (55)	theoretical models data aggregator	http://www.proteinmodelportal.org
TargetTrack (56)	Protein Structure Initiative structural target registration and experimental history database	http://sbkb.org/tt/
KB-Rank Search tool (58)	biomedical annotation structure search	http://protein.tcmedc.org
PSI Technology Portal (59)	catalog of PSI laboratory technologies and methods	http://technology.sbkb.org/portal/
PSI Publications Portal (54)	catalog of articles published by PSI	http://olenka.med.virginia.edu/psi/
Community-Driven Annotation		
Proteopedia (61)	community annotation of structures	http://proteopedia.org
TOPSAN (64)	community annotation of structural genomics structures	http://www.topsan.org