REVIEW

Creating a community resource for protein science

Helen M. Berman*

Department of Chemistry and Chemical Biology, Center for Integrative Proteomics Research, Rutgers, The State University of New Jersey, Piscataway, New Jersey 08854

Received 9 August 2012; Accepted 30 August 2012

DOI: 10.1002/pro.2154

Published online 11 September 2012 proteinscience.org

Helen Berman is the recipient of the Protein Society 2012 Carl Branden Award

Abstract: In addition to being one of the early pioneers in protein crystallography, Carl Brändén made significant contributions to science education with his elegant and beautifully illustrated book *Introduction to Protein Structure* (Brändén and Tooze, New York: Garland, 1991). It is truly an honor to receive this award in their names. This award and the 40th anniversary of the Protein Data Bank (PDB; Berman et al., Structure 2012;20:391–396) have given me an opportunity to reflect on the various components that have contributed to building a resource for protein science and to try to quantify the impact of having PDB data openly available.

Keywords: data resources; community database; structural biology; proteins

Early History of Structural Biology and the Protein Data Bank

With the discovery of Bragg's law in 1912,¹ X-ray crystallography began to be used as a method to determine atomic structure. It did not take long for visionary scientists to explore its use for determining the structure of proteins. In 1934, Bernal and Crowfoot Hodgkin produced the first diffraction pattern of a pepsin crystal.² The determination of the structures of myoglobin^{3,4} and then hemoglobin⁵ earned Perutz and Kendrew Nobel Prizes in 1962. This marked the beginning of an era that has seen extraordinary progress in the use of

X-ray crystallography for structure determination of a wide range of biological molecules for which several more Nobel prizes were awarded.⁶ In the 1990s, nuclear magnetic resonance (NMR) methods began to be exploited for structure determination, and, more recently, 3D electron microscopy has allowed us to visualize the structures of very large molecular machines.

The structures of biological molecules contain a treasure trove of information. There is no doubt that every investigator who determines a structure wants to fully analyze the results of their experiment and probably has the greatest insights into how exactly to do that. At a minimum, these investigators need a place to store their data in a secure space that is preferably not in the local laboratory. But it is also true that others might want to compare, classify, and analyze groups of structures, which would require a way to easily distribute the data. The pioneers of structural biology recognized the necessity for a central repository that could store and distribute structural data, and a group of these scientists

Additional Supporting Information may be found in the online version of this article.

*Correspondence to: Helen M. Berman, Department of Chemistry and Chemical Biology, Center for Integrative Proteomics Research, Rutgers, The State University of New Jersey, Piscataway, New Jersey 08854.
E-mail: berman@rcsb.rutgers.edu.

Grant sponsors: NSF, NIGMS, DOE, NLM, NCI, NINDS, and NIDDK; Grant number: DBI 0829586.

stepped forward to take on the task of creating an archive.⁷ The Protein Data Bank (PDB) was established in 1971 at Brookhaven National Laboratory (BNL) with an initial holding of seven structures.⁸

Components of the PDB

Management

The initial Protein Data Bank (PDB) was managed as collaboration between BNL and the Cambridge Crystallographic Data Centre.⁹ Later, a group in Osaka, Japan, joined the collaboration. All data were annotated at BNL. In 1998, when the Research Collaboratory for Protein Research (RCSB) PDB was awarded the contract from the NSF,¹⁰ a collaboration was established with the PDBj¹¹ group at Osaka University to collect and process data. At the European Bioinformatics Institute in Hinxton, UK, the Macromolecular Structure Database group¹² (now PDBe) also began to collect data. In 2003, the Worldwide PDB (wwPDB) became a formalized collaboration among these three groups who continue to collect and annotate coordinate and related experimental data for the PDB archive.¹³ Later, the BioMagResBank joined as a collection center for NMR spectral and quantitative data-derived data.¹⁴ The purpose of the wwPDB was to ensure that with multiple collection centers, there would be a single global PDB with uniform standards for data processing and validation. A File Transfer Protocol site contains the master archive of data files and is mirrored by the RCSB PDB, PDBe, and PDBj.

Data content

The primary results of a crystal structure determination are the coordinates of every atom in the molecule. For a small protein, there are perhaps 1000 atom sites; for a large one, there are more than 10,000. In a PDB entry, each atom site is identified with atom and residue labels. In addition, there is information about the chemistry of the polymer and small molecule ligands as well as how the structure was determined. For structures determined using X-ray methods, temperature factors and occupancies are included in the atomic site records. Structure factors are also archived, along with restraints and chemical shifts for NMR entries. Electron microscopy (3DEM) entries contain the volume data and the atomic model, where possible.¹⁵

The data deposited into the PDB evolved as structural biology matured. In the case of crystallography, rapid advances in data collection, structure determination, and refinement have necessitated the addition of new data items. Synchrotron sources were not used in crystallographic experiments when the PDB began, and proteins were not routinely refined. With the advent of synchrotron beamlines, new methods for data collection and structure deter-

mination have evolved. Attitudes about what should be collected for the archive have also changed. Structure factors were rarely deposited before 1990. Now structure factor files accompany every X-ray structure. Chemical shift and restraint data are required for NMR structures. As electron microscopy emerges as a powerful method to determine the structures of macromolecular machines, the wwPDB has created data items to describe the models determined by this method and now collects map volumes.

Although the PDB is an archive and could in principle collect the data "as is," PDB entries are very carefully curated in order for the data to be truly useful. Without curation, the PDB would merely be a place for the safe keeping of data for the authors. With curation, users in a variety of fields can more easily view and analyze the data. This enables structural biologists to have more confidence in selecting structures either as a model for molecular replacement or for comparison with their own. Likewise, computational biology has been enabled with the availability of a corpus of data in a uniform format.

How is curation done? Biocurators with backgrounds in crystallography, NMR, and 3DEM review every entry and validate the files with the help of a variety of computer tools. The chemistry and geometry of ligands and polymers are checked against known standards, and structure models are checked against the primary data. In addition, secondary structure, biological assembly, and structure determination and refinement parameters are reviewed. Visualizing the structure to make sure that it looks sensible is also a key part of the curation process. Any unusual findings are reported back to the author, who has an opportunity to revise the file and perhaps rerefine the model.

Infrastructure

As the science of structure determination develops and grows, the PDB needs to be able to adapt and expand. The addition of new data items requires a flexible infrastructure as well as the collaboration of those with expert knowledge of the science.

The PDB began as a collection of flat files in an ASCII format, entered on 80-column punched cards. In time, data on magnetic tape were sent by depositors and then distributed by mail. In 1976, there were 23 structures in the archive, and 375 data sets were distributed to laboratories via magnetic tape. A listing of available entries was distributed with the newsletter; there was not any type of database that could be queried.

The PDB file format, which is based on having 80-characters per line, was established in 1974.⁹ This format has been astonishingly durable and human-readable and is still used by thousands of

software packages. However, it has well-known limitations, most especially with respect to the number of atoms a file can accommodate. About 20 years ago, a decision was made to try and define a format that could accommodate many more atoms, better definitions, and the ability to more easily port the data into a database. The Macromolecular Crystallographic Information File (mmCIF) format and dictionary met these criteria and initially supported more than 3000 data items.¹⁶ As the archive expands to support new methods and richer data descriptions, new data items are added to the dictionary. The name of the format became PDBx and is used internally for data processing and data exchange.¹⁷ Because the legacy format was used so widely, every PDBx file is also converted and made available in the PDB file format. However, this has meant that for large files such as the ribosome, the data are split into multiple files. This does not serve the community well.

About 1 year ago, the wwPDB met with software developers of refinement programs, and it was agreed that the PDBx format should now be the official format for PDB files. These developers are in the process of adjusting their programs to accept and produce PDBx data files. Ample notification to PDB data depositors and users will be required so as to ensure no disruption of service.

The ways in which data have been deposited and annotated have changed over the years. When the RCSB PDB took over the management of the PDB in 1998, it created a new deposition and annotation system called ADIT¹⁰ that could accept either mmCIF or PDB files, but used mmCIF internally. This system was adopted by PDBj. PDBe reengineered BNL's AutoDep system¹⁸ to collect and process data in PDB format.¹⁹ Considerable efforts went into ensuring that although the systems for collecting the data are different, the end product would be the same no matter where the data were processed. In 2007, it was decided that the overhead involved in maintaining these different systems was much too high. The three data collection centers began to collaborate in the creation of a single deposition and annotation system that uses PDBx as the master format and offers significant improvements in processing efficiency and assurance of data quality. As part of this project, the requirements for every aspect of the process were analyzed using the combined experience in data processing of all the centers. The system will begin testing in Fall 2012.

Another key wwPDB collaboration is the regular review and remediation of the data across the archive. These reviews become particularly important as new methods and discovery begin to challenge how all structures are curated and represented in the archive. In the past reviews, formats of atoms names have been made uniform, details

about polymer and ligand chemistry have been improved,²⁰ and the representation of antibiotics and peptide inhibitors has been made uniform. Viruses are now represented in more useful ways.²¹ The next project we are tackling is the representation of carbohydrates in the archive.

The wwPDB centers are ever mindful that the community expects 24/7 operations and the rapid delivery of quality data. We also know that although users may complain about some aspects of the data, they are also hesitant about change that could cause the disruption of PDB services or to the software they use in their labs. So, the challenge is to continue to improve the underlying data without causing problems for the users. Thus, a considerable amount of planning must go into any new rollout of data.

Although the wwPDB data centers collaborate on annotating data, so that the PDB files are the same no matter where they are retrieved, each site provides different services for analyzing, visualizing, and comparing the data.^{11,22–24} These multiple sites, tools, and ways of exploring the data are incredibly useful for “exercising” the data.

Community stakeholders

By definition, a community database must engage with its users. From Day 1, the PDB needed the trust and support of the structure authors. In the early days, this meant convincing researchers to deposit their data and working with them to be sure the data are represented accurately. A few very active protein crystallographers, among them Fred Richards, Max Perutz, and Michael Rossmann, were vocal in their support of this enterprise. Although many people did voluntarily deposit data, others chose to keep the data private. In 1982, the very visionary NIH program officer Marvin Cassman became concerned that data key to the development of drugs for the emerging public health threat of AIDS might not see its way into the PDB. Richards and Dick Dickerson were very vocal in their concerns that valuable data would be lost if data deposition was not mandatory.²⁵ They circulated a petition urging mandatory data deposition, signed by 182 prominent scientists from around the globe. The International Union of Crystallography established a committee charged with creating deposition guidelines. The committee spent a great deal of time defining the scope of what kinds of data should be deposited and when. The guidelines were published in 1989.²⁶ These guidelines then became the basis for journal and funding policies. Today, virtually every journal requires data deposition for publication. It is very important to note here that the data providers who know their data best were the ones who came up with sensible and enforceable guidelines.

Figure 1. Growth of the PDB archive: (a) deposited (in black) and released (gray) structures shown using a logarithmic scale; (b) total number of unique nonpolymer ligands released each year (a single entry may have several ligands); (c) number of peptide-like inhibitor/antibiotic entries released per year. There are three notable peaks in (c): 73 structures with an inhibitor/antibiotic were released in 1994, the majority of which are thrombin inhibitors and renin inhibitors; 130 structures in 2006, the majority of which are thrombin inhibitors and protease inhibitors; and 140 structures in 2011, the majority of which are protease inhibitors and caspase inhibitors.

The same process is being used for creating validation standards. The wwPDB has set up a group of Task Forces who have been charged with coming up with best practices for validating data. So far, Validation Task Forces (VTF) have been set up in X-ray crystallography, NMR spectroscopy, electron microscopy,²⁷ and now small-angle scattering. The X-ray VTF published a detailed set of recommendations based on careful analysis of the data.²⁸ These recommendations are now being implemented by the wwPDB and will be part of the new Deposition and Annotation tool. The other task forces will follow

suit. The extraordinary care by which these validation standards are being set by experts in the field will ensure that the quality of the data in the archive will continue to be high.

The journals are also key players. Just as every published structure must be deposited in the PDB, journals are beginning to require the submission of validation reports as part of the editorial process as a result of the work of the VTFs.^{29,30}

Most people who use the PDB are not experimental structural biologists. Computational biologists use the data for classification, analysis, and structure prediction. Experimental biologists use the data to better understand protein function. Educators from grade school to postgraduate level use the PDB in their classrooms. For these users, the wwPDB websites provide rich features and tools such as interactive graphics and educational content. The very widespread use of the data makes it particularly important that it is well curated and that there are presentations of the data that are accessible to a diverse audience.

Trends in the data

The PDB has grown from an archive of seven structures with relatively low-molecular weight to now more than 83,000 structures (Fig. 1). At the current

Figure 2. Growth in the size (assessed by molecular weight) of released structures in the PDB for entries determined by X-ray crystallography (gray) and NMR (black). For X-ray, the molecular mass was calculated without solvent, corrected for noncrystallographic symmetry (NCS), and combining split entries into single structures. For viruses or entries that used noncrystallographic symmetry, molecular weights for the entire asymmetric unit were calculated by multiplying the molecular weight of the polymeric chains by the number of nonidentity NCS operators. For NMR, mass was determined excluding water. The large increase shown in 1984 is due to the release of the tomato bushy stunt virus 2tbv.³¹

Figure 3. Growth in the use of synchrotron X-ray sources in structure determination as determined from X-ray source annotation. The number of structures determined using synchrotron radiation is shown in gray and the number using home-laboratory sources in black. This plot shows that while the use of home sources in X-ray structure determination has remained roughly constant, the use of synchrotron sources has increased rapidly.

rate of deposition, it is projected that 10,000 structures will be deposited in 2012—more than the size of the entire archive in 2000. Will the growth rate continue? Yes, but perhaps not at the rate seen in the 1990s. As shown in Figure 2, the average molecular weight deposited per year has more than tripled since the PDB began. This is of course due to the vast improvements in technology in all structure-determination methods. For example, synchrotron sources are now used routinely for X-ray data

Figure 4. Range of the reported high-resolution limit in released X-ray structures in the PDB archive. The resolution limit range of all entries is indicated by the green area. The blue-dotted line indicates the limit of the range if the outlying lowest-resolution structure for that year is removed. The mean resolution limit for a given year is indicated by the dashed red line. The median value after 1989 (not shown) is roughly the same as the mean. Methods developed in the last decade can handle the refinement of lower resolution structures, the value of which has been accepted by the community.

Table I. Clusters of proteins with redundant sequences

Cluster number	Number of structures	Protein cluster	Biological connection
1	522, 374, 203	Lysozyme: bacteriophage T4, hen egg white, human	Evolution, folding, catalysis
2	446	HIV-1 protease	Human immunodeficiency virus (HIV)
3	394	Human carbonic anhydrase II	Osteopetrosis
4	379, 300, 266	Trypsin, thrombin, thrombin light chain	Catalysis, blood clotting
5	361	β -2 microglobulin	MHC complex, virus protection, cancer
6	229, 199, 197	Whale myoglobin, human hemoglobin β -subunit, human hemoglobin α -subunit	Evolution, sickle-cell anemia, thalassemia
7	217	CDK2	DNA damage repair, cancer, cell proliferation
8	190	Ribonuclease A	Specificity, catalysis
9	188	MAP kinase 14	Arthritis, stresses, and proinflammatory cytokines
10	188	β -Secretase 1	Alzheimer's disease
11	171, 169	Insulin A, B chain	Diabetes
12	166, 159	Reverse-transcriptase RNase H, p51	Human immunodeficiency virus (HIV)
13	162, 158	Class I histocompatibility antigen A-2 (human, mouse)	Antibody recognition
14	157	Glycogen phosphorylase	Type 2 diabetes, catalysis
15	146+	Ribosomal proteins	Protein biosynthesis
16	154	Cytochrome C peroxidase	Oxidative chemistry
17	152	Transthyretin	Amyloid diseases
18	149	Green fluorescent protein	Bioluminescence

Using the sequence data of all protein chains (as of May 2012), a clustering of sequences was created using BLASTclust³⁵ with a 95% sequence identity cutoff. Resulting clusters were then ranked based on the number of structures and tabulated. Differing clusters of proteins with related functions are reported on the same line separated by commas. The PDB IDs of the entries in each cluster are available in the supplementary materials.

Figure 5. HIV and drugs bound to HIV protease and reverse transcriptase. (a) HIV structures in the PDB.^{37–53} Image by David Goodsell is available from the RCSB PDB website. (b) HIV-1 reverse transcriptase,⁵⁴ HIV-1 protease,⁵⁵ and related FDA-approved inhibitors. Inhibitors shown in green are available in the PDB bound to a protein and labeled with the first year an example was released in the PDB. Inhibitors shown in gray have not been deposited in the PDB. As of July 2012, there are more than 160 structures of HIV-1 reverse transcriptase in the PDB; ~ 75 are bound to a drug or inhibitor. There are almost 450 structures of HIV-1 protease in the PDB; ~ 260 are bound to a drug or inhibitor.

collection (Fig. 3). Scientists have been enabled to take on increasingly challenging problems to answer questions in biology, and they do. As part of the NIH's PSI:Biology program (Pieper et al., in preparation),³² nine groups are tackling notoriously difficult membrane proteins. As we develop the methods to determine these structures, the rate of structure production may decrease before we see the expected increase as we did when high-throughput methods were perfected. A slowdown occurred in 2008, when 7073 entries were deposited; 8130 entries were deposited in the previous year. This change correlates with the discontinuation of the Structural Genomics group in Japan in 2007.³³ Similar abrupt changes in funding of large projects and facilities anywhere in the world could have the same effect. Another trend is the increase in the number of ligand bound structures deposited each year [Fig. 1(b)]. These ligands range from very simple organic molecules to complex peptide-like inhibitors and antibiotics [Fig. 1(c)]. The latter increase is attributed to the attempts to design inhibitors for proteases, thrombin, and, more recently, the ribosome. The data contributed by the

public sector are no doubt only a fraction of the data being produced by the pharmaceutical industry, most of which are not archived in the PDB.

Although the average resolution of X-ray structures in the PDB has stayed at 2.1 Å since 1987, a substantial number are better than that average, thus yielding very high-quality structures (Fig. 4). The number of low-resolution structures has also increased and can be correlated with the increased activity focused on molecular machines of very high-molecular weight.

The effect of doing routine validation of structures before deposition into the PDB has been demonstrated to have a dramatic effect on the quality of structures treated in that way. During PSI efforts 1 and 2, all structures were thoroughly checked by the structural genomics centers before deposition, and very high-quality structures were produced.³⁴ The initial fear that high-throughput would result in inferior structures was unfounded. Implementing the recommendations of the wwPDB VTF will help ensure that all structural biologists will perform validation routinely.

b

Reverse Transcriptase Inhibitors

Nucleoside Reverse Transcriptase Inhibitors (NRTIs)

Non-nucleoside Reverse Transcriptase Inhibitors (NNRTIs)

Protease Inhibitors (PIs)

Figure 5. (Continued).

In the search for new folds, it is important to select structures containing sequences with much less than 30% sequence similarity. And even with that criteria, there is no guarantee that the protein will have a new fold, as was demonstrated in the early phases of the PSI structural genomics efforts. To understand protein function, however, working with redundant sequences is crucial (Table I). Work on phage T4 lysozyme, in which structures contain-

ing sequences with systematic substitutions were determined, gave us dramatic insights into protein folding.³⁶ The determination of the structures of HIV protease and reverse transcriptase bound to different ligands was critical to the development of effective drugs against AIDS (Fig. 5).^{56,57}

The PDB can now be used to describe whole systems in biology. For example, it has been possible to describe the structures of the proteins expressed in

Figure 6. The tricarboxylic acid cycle, also known as the Krebs cycle, shown from a structural biology perspective. Shown are the images of full (or nearly full) structures available in the PDB for all of the enzymes in the cycle.^{60–69}

a whole organism (*Thermotoga maritima*),⁵⁸ and there has been significant progress in our understanding of proteins involved in cancer.⁵⁹ Significantly, it is now possible to have a molecular description of entire pathways. Although it took 25 years, we now have structures of all the enzymes involved in the tricarboxylic acid cycle (Fig. 6). Similar results can be seen with respect to other pathways.⁷⁰

The seemingly impossible dream of having a structural view of biology is much closer than we thought.

Acknowledgments

Many people at the wwPDB data centers support the continued viability and quality of the PDB archive, in particular that of the leaders of the member sites Kim Henrick, Gerard Kleywegt, Haruki Nakamura, and John Markley. The work of RCSB PDB team at Rutgers University and the University of California San Diego over the years is greatly appreciated, especially that of John Westbrook, Phil Bourne, Zukang Feng, and Christine Zardecki. Special thanks to Brian Hudson and Ezra Peisach for their thoughtful help

with this manuscript. Helen Berman is the recipient of the Protein Society 2012 Carl Brändén Award.

References

1. Bragg WL (1913) The diffraction of short electromagnetic waves by a crystal. *Proc Camb Phil Soc* 17:43–57.
2. Bernal JD, Crowfoot DM (1934) X-ray photographs of crystalline pepsin. *Nature* 133:794–795.
3. Kendrew JC, Bodo G, Dintzis HM, Parrish RG, Wyckoff H, Phillips DC (1958) A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. *Nature* 181:662–666.
4. Kendrew JC, Dickerson RE, Strandberg BE, Hart RG, Davies DR, Phillips DC, Shore VC (1960) Structure of myoglobin: a three-dimensional Fourier synthesis at 2 Å resolution. *Nature* 185:422–427.
5. Perutz MF, Rossmann MG, Cullis AF, Muirhead H, Will G, North ACT (1960) Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5 Å resolution, obtained by X-ray analysis. *Nature* 185:416–422.
6. International Union of Crystallography (2011) Nobel Prize winners associated with crystallography. <http://www.iucr.org/people/nobel-prize>.
7. Berman H (2008) The Protein Data Bank: a historical perspective. *Acta Cryst A* 64:88–95.
8. Protein Data Bank (1971) Protein Data Bank. *Nature New Biol* 233:223.

9. Bernstein FC, Koetzle TF, Williams GJB, Meyer EF, Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M (1977) Protein Data Bank: a computer-based archival file for macromolecular structures. *J Mol Biol* 112:535–542.
10. Berman HM, Westbrook JD, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. *Nucleic Acids Res* 28:235–242.
11. Kinjo AR, Suzuki H, Yamashita R, Ikegawa Y, Kudou T, Igarashi R, Kengaku Y, Cho H, Standley DM, Nakagawa A, Nakamura H (2012) Protein Data Bank Japan (PDBj): maintaining a structural data archive and resource description framework format. *Nucleic Acids Res* 40:D453–D460.
12. Boutselakis H, Dimitropoulos D, Fillon J, Golovin A, Henrick K, Hussain A, Ionides J, John M, Keller PA, Krissinel E, McNeil P, Naim A, Newman R, Oldfield T, Pineda J, Rachedi A, Copeland J, Sitnov A, Sobhany S, Suarez-Uruena A, Swaminathan J, Tagari M, Tate J, Tromm S, Velankar S, Vranken W (2003) E-MSD: the European Bioinformatics Institute Macromolecular Structure Database. *Nucleic Acids Res* 31:458–462.
13. Berman HM, Henrick K, Nakamura H (2003) Announcing the worldwide Protein Data Bank. *Nat Struct Biol* 10:980.
14. Markley JL, Ulrich EL, Berman HM, Henrick K, Nakamura H, Akutsu H (2008) BioMagResBank (BMRB) as a partner in the Worldwide Protein Data Bank (wwPDB): new policies affecting biomolecular NMR depositions. *J Biomol NMR* 40:153–155.
15. Lawson CL, Baker ML, Best C, Bi C, Dougherty M, Feng P, van Ginkel G, Devkota B, Lagerstedt I, Ludtke SJ, Newman RH, Oldfield TJ, Rees I, Sahni G, Sala R, Velankar S, Warren J, Westbrook JD, Henrick K, Kleywegt GJ, Berman HM, Chiu W (2011) EMDatabank.org: unified data resource for CryoEM. *Nucleic Acids Res* 39:D456–D464.
16. Fitzgerald PMD, Westbrook JD, Bourne PE, McMahon B, Watenpaugh KD, Berman HM, 4.5 Macromolecular dictionary (mmCIF). In: Hall SR, McMahon B, Eds. (2005) *International tables for crystallography*, Vol. G: definition and exchange of crystallographic data. Springer: Dordrecht, The Netherlands, pp. 295–443.
17. Westbrook J, Henrick K, Ulrich EL, Berman HM, 3.6.2 The Protein Data Bank exchange data dictionary. In Hall SR, McMahon B, Eds. (2005) *International tables for crystallography*, Vol. G: definition and exchange of crystallographic data. Springer: Dordrecht, The Netherlands, pp. 195–198.
18. Sussman JL, Lin D, Jiang J, Manning NO, Prilusky J, Ritter O, Abola EE (1998) Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules. *Acta Cryst D* 54:1078–1084.
19. Keller PA, Henrick K, McNeil P, Moodie S, Barton GJ (1998) Deposition of macromolecular structures. *Acta Cryst D* 54:1105–1108.
20. Henrick K, Feng Z, Bluhm WF, Dimitropoulos D, Doreleijers JF, Dutta S, Flippen-Anderson JL, Ionides J, Kamada C, Krissinel E, Lawson CL, Markley JL, Nakamura H, Newman R, Shimizu Y, Swaminathan J, Velankar S, Ory J, Ulrich EL, Vranken W, Westbrook J, Yamashita R, Yang H, Young J, Yousufuddin M, Berman HM (2008) Remediation of the Protein Data Bank archive. *Nucleic Acids Res* 36:D426–D433.
21. Lawson CL, Dutta S, Westbrook JD, Henrick K, Berman HM (2008) Representation of viruses in the remediated PDB archive. *Acta Cryst D* 64:874–882.
22. Rose PW, Beran B, Bi C, Bluhm WF, Dimitropoulos D, Goodsell DS, Prlic A, Quesada M, Quinn GB, Westbrook JD, Young J, Yukich B, Zardecki C, Berman HM, Bourne PE (2011) The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Res* 39:D392–D401.
23. Velankar S, Alhroub Y, Best C, Caboche S, Conroy MJ, Dana JM, Fernandez Montecelo MA, van Ginkel G, Golovin A, Gore SP, Gutmanas A, Haslam P, Hendrickx PM, Heuson E, Hirshberg M, John M, Lagerstedt I, Mir S, Newman LE, Oldfield TJ, Patwardhan A, Rinaldi L, Sahni G, Sanz-Garcia E, Sen S, Slowley R, Suarez-Uruena A, Swaminathan GJ, Symmons MF, Vranken WF, Wainwright M, Kleywegt GJ (2012) PDBe: Protein Data Bank in Europe. *Nucleic Acids Res* 40:D445–D452.
24. Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, Livny M, Mading S, Maziuk D, Miller Z, Nakatani E, Schulte CF, Tolmie DE, Wenger RK, Yao H, Markley JL (2008) BioMagResBank. *Nucleic Acids Res* 36:D402–D408.
25. Drew HR, Samson S, Dickerson RE (1982) Structure of a B-DNA dodecamer at 16K. *Proc Natl Acad Sci USA* 79:4040–4044.
26. International Union of Crystallography (1989) Policy on publication and the deposition of data from crystallographic studies of biological macromolecules. *Acta Cryst A* 45:658.
27. Henderson R, Sali A, Baker ML, Carragher B, Devkota B, Downing KH, Egelman EH, Feng Z, Frank J, Grigorieff N, Jiang W, Ludtke SJ, Medalia O, Penczek PA, Rosenthal PB, Rossmann MG, Schmid MF, Schroder GF, Steven AC, Stokes DL, Westbrook JD, Wriggers W, Yang H, Young J, Berman HM, Chiu W, Kleywegt GJ, Lawson CL (2012) Outcome of the first electron microscopy validation task force meeting. *Structure* 20:205–214.
28. Read RJ, Adams PD, Arendall WB, III, Brunger AT, Emsley P, Joosten RP, Kleywegt GJ, Krissinel EB, Lutke T, Otwinowski Z, Perrakis A, Richardson JS, Sheffler WH, Smith JL, Tickle IJ, Vriend G, Zwart PH (2011) A new generation of crystallographic validation tools for the Protein Data Bank. *Structure* 19:1395–1412.
29. Baker EN, Dauter Z, Einspahr H, Weiss MS (2010) In defence of our science—validation now! *Acta Cryst D* 66:115.
30. Allewell NM (2012) Validating macromolecular structures. *J Biochem* 287:2295.
31. Hopper P, Harrison SC, Sauer RT (1984) Structure of tomato bushy stunt virus. V. Coat protein sequence determination and its structural implications. *J Mol Biol* 177:701–713.
32. National Institute of General Medical Sciences (2010) PSI: Biology Network. http://www.nigms.nih.gov/Research/FeaturedPrograms/PSI/psi_biology/psibiology_networkorg.htm.
33. Protein 3000 project over—aim achieved (2007) *Riken Res* 2:17.
34. Chruszcz M, Domagalski M, Osinski T, Wlodawer A, Minor W (2010) Unmet challenges of structural genomics. *Curr Opin Struct Biol* 20:587–597.
35. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215:403–410.
36. Baase WA, Liu L, Tronrud DE, Matthews BW (2010) Lessons from the lysozyme of phage T4. *Protein Sci* 19: 631–641.
37. Hill CP, Worthylake D, Bancroft DP, Christensen AM, Sundquist WI (1996) Crystal structures of the trimeric human immunodeficiency virus type 1 matrix protein: implications for membrane association and assembly. *Proc Natl Acad Sci USA* 93:3099–3104.
38. Pornillos O, Ganser-Pornillos BK, Kelly BN, Hua Y, Whitby FG, Stout CD, Sundquist WI, Hill CP, Yeager

- M (2009) X-ray structures of the hexameric building block of the HIV capsid. *Cell* 137:1282–1292.
39. Kwong PD, Wyatt R, Majeed S, Robinson J, Sweet RW, Sodroski J, Hendrickson WA (2000) Structures of HIV-1 gp120 envelope glycoproteins from laboratory-adapted and primary isolates. *Structure* 8:1329–1339.
 40. Caffrey M, Cai M, Kaufman J, Stahl SJ, Wingfield PT, Covell DG, Gronenborn AM, Clore GM (1998) Three-dimensional solution structure of the 44 kDa ectodomain of SIV gp41. *EMBO J* 17:4572–4584.
 41. De Guzman RN, Wu ZR, Stalling CC, Pappalardo L, Borer PN, Summers MF (1998) Structure of the HIV-1 nucleocapsid protein bound to the SL3 psi-RNA recognition element. *Science* 279:384–388.
 42. Ye X, Kumar RA, Patel DJ (1995) Molecular recognition in the bovine immunodeficiency virus Tat peptide-TAR RNA complex. *Chem Biol* 2:827–840.
 43. Peloponese JM, Jr, Gregoire C, Opi S, Esquieu D, Sturgis J, Lebrun E, Meurs E, Collette Y, Olive D, Aubertin AM, Witvrow M, Pannecouque C, De Clercq E, Bailly C, Lebreton J, Loret EP (2000) ^1H - ^{13}C nuclear magnetic resonance assignment and structural characterization of HIV-1 Tat protein. *C R Acad Sci III* 323:883–894.
 44. Battiste JL, Mao H, Rao NS, Tan R, Muhandiram DR, Kay LE, Frankel AD, Williamson JR (1996) Alpha helix-RNA major groove recognition in an HIV-1 rev peptide-RRE RNA complex. *Science* 273:1547–1551.
 45. Arold S, Franken P, Strub MP, Hoh F, Benichou S, Benarous R, Dumas C (1997) The crystal structure of HIV-1 Nef protein bound to the Fyn kinase SH3 domain suggests a role for this complex in altered T cell receptor signaling. *Structure* 5:1361–1372.
 46. Geyer M, Munte CE, Schorr J, Kellner R, Kalbitzer HR (1999) Structure of the anchor-domain of myristoylated and non-myristoylated HIV-1 Nef protein. *J Mol Biol* 289:123–138.
 47. Wecker K, Morellet N, Bouaziz S, Roques BP (2002) NMR structure of the HIV-1 regulatory protein Vpr in H_2O /trifluoroethanol. Comparison with the Vpr N-terminal (1–51) and C-terminal (52–96) domains. *Eur J Biochem* 269:3779–3788.
 48. Stanley BJ, Ehrlich ES, Short L, Yu Y, Xiao Z, Yu XF, Xiong Y (2008) Structural insight into the human immunodeficiency virus Vif SOCS box and its role in human E3 ubiquitin ligase assembly. *J Virol* 82:8656–8663.
 49. Park SH, Mrse AA, Nevzorov AA, Mesleh MF, Oblatt-Montal M, Montal M, Opella SJ (2003) Three-dimensional structure of the channel-forming trans-membrane domain of virus protein “u” (Vpu) from HIV-1. *J Mol Biol* 333:409–424.
 50. Willbold D, Hoffmann S, Rosch P (1997) Secondary structure and tertiary fold of the human immunodeficiency virus protein U (Vpu) cytoplasmic domain in solution. *Eur J Biochem* 245:581–588.
 51. Sarafianos SG, Das K, Tantillo C, Clark AD, Jr, Ding J, Whitcomb J, Boyer PL, Hughes SH, Arnold E (2001) Crystal structure of HIV-1 reverse transcriptase in complex with a polypurine tract RNA:DNA. *EMBO J* 20:1449–1461.
 52. Chen JC, Krucinski J, Miercke LJ, Finer-Moore JS, Tang AH, Leavitt AD, Stroud RM (2000) Crystal structure of the HIV-1 integrase catalytic core and C-terminal domains: a model for viral DNA binding. *Proc Natl Acad Sci USA* 97:8233–8238.
 53. Kim EE, Baker CT, Dwyer MD, Mureko MA, Rao BG, Tung RD, Navia MA (1995) Crystal structure of HIV-1 protease in complex with Vx-478, a potent and orally bioavailable inhibitor of the enzyme. *J Am Chem Soc* 117:1181–1182.
 54. Smerdon SJ, Jager J, Wang J, Kohlstaedt LA, Chirino AJ, Friedman JM, Rice PA, Steitz TA (1994) Structure of the binding site for nonnucleoside inhibitors of the reverse transcriptase of human immunodeficiency virus type 1. *Proc Natl Acad Sci USA* 91:3911–3915.
 55. Wlodawer A, Miller M, Jaskolski M, Sathyanarayana BK, Baldwin E, Weber IT, Selk LM, Clawson L, Schneider J, Kent SB (1989) Conserved folding in retroviral proteases: crystal structure of a synthetic HIV-1 protease. *Science* 245:616–621.
 56. Wensing AM, van Maarseveen NM, Nijhuis M (2010) Fifteen years of HIV protease inhibitors: raising the barrier to resistance. *Antiviral Res* 85:59–74.
 57. Cihlar T, Ray AS (2010) Nucleoside and nucleotide HIV reverse transcriptase inhibitors: 25 years after zidovudine. *Antiviral Res* 85:39–58.
 58. Zhang Y, Thiele I, Weekes D, Li Z, Jaroszewski L, Ginalski K, Deacon AM, Wooley J, Lesley SA, Wilson IA, Pals-son B, Osterman A, Godzik A (2009) Three-dimensional structural view of the central metabolic network of *Thermotoga maritima*. *Science* 325:1544–1549.
 59. Huang YJ, Hang D, Lu LJ, Tong L, Gerstein MB, Montelione GT (2008) Targeting the human cancer pathway protein interaction network by structural genomics. *Mol Cell Proteom* 7:2048–2060.
 60. Birktoft JJ, Rhodes G, Banaszak LJ (1989) Refined crystal structure of cytoplasmic malate dehydrogenase at 2.5-Å resolution. *Biochemistry* 28:6065–6081.
 61. Remington S, Wiegand G, Huber R (1982) Crystallographic refinement and atomic models of two different forms of citrate synthase at 2.7 and 1.7 Å resolution. *J Mol Biol* 158:111–152.
 62. Robbins AH, Stout CD (1989) Structure of activated aconitase: formation of the [4Fe-4S] cluster in the crystal. *Proc Natl Acad Sci USA* 86:3639–3643.
 63. Hurley JH, Thorsness PE, Ramalingam V, Helmers NH, Koshland DE, Jr, Stroud RM (1989) Structure of a bacterial enzyme regulated by phosphorylation, isocitrate dehydrogenase. *Proc Natl Acad Sci USA* 86:8635–8639.
 64. Frank RA, Price AJ, Northrop FD, Perham RN, Luisi BF (2007) Crystal structure of the E1 component of the *Escherichia coli* 2-oxoglutarate dehydrogenase multienzyme complex. *J Mol Biol* 368:639–651.
 65. Knapp JE, Mitchell DT, Yazdi MA, Ernst SR, Reed LJ, Hackert ML (1998) Crystal structure of the truncated cubic core component of the *Escherichia coli* 2-oxoglutarate dehydrogenase multienzyme complex. *J Mol Biol* 280:655–668.
 66. Mattevi A, Obmolova G, Kalk KH, van Berkel WJ, Hol WG (1993) Three-dimensional structure of lipoamide dehydrogenase from *Pseudomonas fluorescens* at 2.8 Å resolution. Analysis of redox and thermostability properties. *J Mol Biol* 230:1200–1215.
 67. Wolodko WT, Fraser ME, James MN, Bridger WA (1994) The crystal structure of succinyl-CoA synthetase from *Escherichia coli* at 2.5-Å resolution. *J Biol Chem* 269:10883–10890.
 68. Yankovskaya V, Horsefield R, Tornroth S, Luna-Chavez C, Miyoshi H, Leger C, Byrne B, Cecchini G, Iwata S (2003) Architecture of succinate dehydrogenase and reactive oxygen species generation. *Science* 299:700–704.
 69. Weaver T, Banaszak L (1996) Crystallographic studies of the catalytic and a second site in fumarate C from *Escherichia coli*. *Biochemistry* 35:13955–13965.
 70. Julfayev ES, McLaughlin RJ, Tao YP, McLaughlin WA (2012) KB-Rank: efficient protein structure and functional annotation identification via text query. *J Struct Funct Genomics* 13:101–110.