Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 1 June 2007
Accepted 20 July 2007

The Protein Data Bank: a historical perspective

Helen M. Berman

Rutgers, The State University of New Jersey, USA. Correspondence e-mail: berman@rcsb.rutgers.edu

The Protein Data Bank began as a grassroots effort in 1971. It has grown from a small archive containing a dozen structures to a major international resource for structural biology containing more than 40000 entries. The interplay of science, technology and attitudes about data sharing have all played a role in the growth of this resource.

1. The history of the Protein Data Bank

The establishment of the Protein Data Bank (PDB) began in the 1970's as a grassroots effort. A group of (then) young crystallographers, including Edgar Meyer, Gerson Cohen and myself, began discussing the idea of establishing a central repository for coordinate data at an American Crystallographic Association (ACA) meeting in Ottawa, Canada, in 1970. Those conversations were continued with a larger group at the ACA meeting in Columbia, South Carolina, USA, in 1971. At that meeting, a petition was written and a proposal was submitted to the United States National Committee for Crystallography (USNCCr). Later that year, the Cold Spring Harbor (CSH) Symposium was held on 'Structure and Function of Proteins at the Three Dimensional Level' (Cold Spring Laboratory Press, 1972). This meeting, characterized by David Phillips as a 'coming of age', heralded a new era in biology. The discussions within the meeting room, on the lawn, and on the beach were exciting and intense. In an informal meeting convened by Max Perutz, protein crystallographers discussed how best to collect and distribute data. Until that point, coordinates for individual entries had only been exchanged among a few research laboratories using punched cards. Since each atom was represented by a single card, an exchange of a structure the size of myoglobin required more than 1000 cards. By providing a central repository for these data, the PDB would make such an exchange possible for anyone.

Walter Hamilton was also in attendance. A chemist at Brookhaven National Laboratory (BNL) and a leader in the crystallographic community, Hamilton had begun to focus on two new science and technology projects. In collaboration with postdoctoral fellow Tom Koetzle and others, he was working on the determination of the structures of all the amino acids using neutron diffraction (Lehmann *et al.*, 1972). In another collaboration, he was developing new computer technologies for graphics and for remote computing with Edgar Meyer (Meyer, 1997). During the CSH meeting, Hamilton was approached with the idea that had been discussed within the ACA community – a public data bank of protein structures. At an *ad hoc* meeting of protein crystallographers attending the Symposium, it was proposed that there should be a repository with identical files in the United Kingdom and in the USA. Hamilton volunteered to set up the American data bank at Brookhaven.

When Max Perutz returned to England, he discussed this proposal with Olga Kennard, who was the founder of the Cambridge Crystallographic Data Centre (CCDC) (Kennard *et al.*, 1972; Allen *et al.*, 1973), and had wide experience in assembling and archiving crystallographic data. Walter Hamilton wrote to her with an offer of collaboration and proposed to meet and discuss some of the details of coordinating the activities. He visited England that summer and, by October 1971, the establishment of the Protein Data Bank archive, jointly operated by the CCDC and BNL, was announced in *Nature New Biology* (Protein Data Bank, 1971). After Hamilton's untimely death in 1973, Koetzle took over the direction of the PDB, and with the support of key members of the community – most especially Michael Rossmann and Fred Richards – the PDB was able to survive. In 1974, the first *PDB Newsletter* was distributed to describe the details of data deposition and remote access. At this point, thirteen structures were ready for distribution and four were pending.

According to the January 1976 report to the ACA Council, the PDB archive contained 23 structures and 375 data sets had been distributed to 31 laboratories in that year. A grant for USD 33000 from the National Science Foundation was awarded, and an Advisory Board consisting of David Davies, Fred Richards and Ken Neet had been established. The project, which began as a dream of a community, finally had all the components of a fully fledged international resource (Bernstein *et al.*, 1977).

The PDB remained in Brookhaven until 1998. In 1999, the management changed to a consortium called the Research Collaboratory of Structural Bioinformatics (RCSB PDB) consisting of Rutgers, The State University of New Jersey, the San Diego Supercomputer Center at the University of California San Diego (UCSD) and the National Institute of Standards and Technology (Berman *et al.*, 2000). In 2003, the Worldwide PDB (wwPDB) formalized the existing international collaborations and an agreement was made among the RCSB PDB, the Macromolecular Structure Database at the

European Bioinformatics Institute (MSD-EBI) and PDB Japan (PDBj) at the Institute for Protein Research at Osaka University, Japan (Berman *et al.*, 2003). This set the stage for the PDB to remain a single uniform global resource of structural biology data.

2. Impact of technology and community action

The first decade of the PDB was characterized by attempts to capture the interest of the community in depositing their structures. Because there were relatively few structural biology groups at that time, it was possible to know virtually every member of the structural biology community. Letters were written and phone calls made to authors of articles reporting structures requesting that they deposit their coordinates in the PDB archive.

The rate of structure determinations began to change as technology became more advanced. The 1980's saw rapid development in all aspects of the structure determination pipeline. Molecular biology made it possible to clone genes and express proteins. Crystallization methods began to evolve and sparse-matrix methods were introduced (Jancarik & Kim, 1991). Synchrotron beamlines made it possible to obtain extremely intense X-rays (Helliwell, 1983). Detection methods evolved from collecting one reflection at a time to multiwire detectors to CCDs. Computer speeds and storage capacities increased dramatically. Graphics computers made it possible to visualize molecules and electron density. New computational methods for all stages of structure determination were developed. Multiple anomalous dispersion (MAD) using synchrotron radiation enabled the direct determination of structures (Hendrickson, 1991). Programs such as *FRODO* (Jones, 1978) and *O* (Jones *et al.*, 1991) allowed computerized electron-density fitting. Methods to refine structures evolved with the use of geometrical and energy restraints (Hendrickson & Konnert, 1981; Brünger, 1990).

As it became easier to determine structures, more and more structures were published. Soon it became commonplace to see pictures of structures on the covers of journals such as *Nature* and *Science*. New journals, including *Acta*

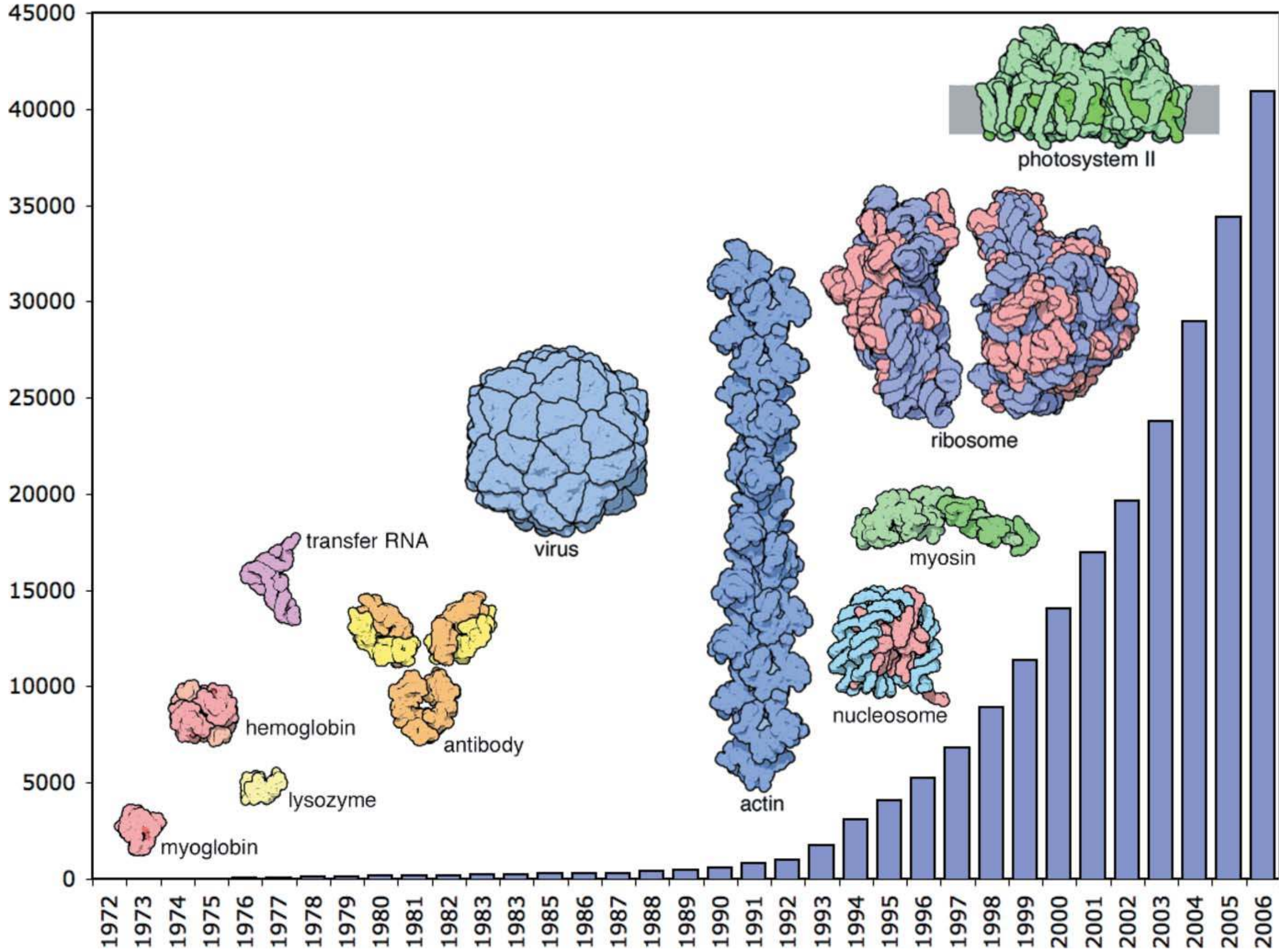

Figure 1
The growth of the number of structures in the PDB archive 1972–2006.

Crystallographica Section D and *Nature Structural Biology* (now *Nature Structural & Molecular Biology*), were established to report the results of the analyses of biological macromolecules. Many now believed that structural biology could give definitive information that would be key to understanding the molecular basis of biology and medicine.

As the value of structural biology became more obvious to other biologists, several committees were formed to look at the new demands from the community for required data deposition. The IUCr Commission on Biological Macromolecules set up a committee to determine exact policies. The ACA and the USNCCr also set up a committee. Fred Richards created an *ad hoc* group of crystallographers who felt strongly about creating a policy that would require data deposition. These groups worked for a few years to hammer out the exact guidelines. The timing of deposition of coordinates and of structure factors was discussed at length. During this period, commentaries were published in journals, letters to the editor were written and there was intense debate in the community. By the end of the decade, a petition resulting from the action of the Richards committee was produced and, in 1989, a formal guideline for data deposition was published (International Union of Crystallography, 1989). The guidelines stated that coordinates should be deposited at the time of publication and released within one year; structure factors should be deposited and released within four years. Most major journals adopted these guidelines and the National Institute for General Medical Sciences made the bold step of saying that funding would be contingent on the open sharing of structural data.

3. The content of the PDB

By the end of the 1980's, the number of structures in the PDB began to increase dramatically (Fig. 1) and that growth continues to date. Nuclear magnetic resonance (NMR) methods began to be used to determine structures, thus providing additional types of information to be archived in the PDB. These structures now make up about 15% of the structures in the archive.

In addition, the complexity of the structures that could be determined grew. The use of flash freezing and highly intense synchrotron radiation during data collection combined with the sophisticated use of non-crystallographic symmetry for structure determination made it possible to solve virus structures (Arnold & Rossmann, 1988) (Fig. 2). Even larger structures, including molecular machines such as the ribosome particles, became amenable to the methods of X-ray crystallography (Moore, 2001). The ability to freeze single particles also allowed structures to be determined using cryoelectron microscopy (cryoEM) and models for these structures began to appear in the PDB archive (Fig. 3).

By the year 2000, the diversity of molecules and complexes in the PDB archive (Fig. 4) was so great that the possibility of actually understanding biology and medicine at a molecular level was not just a far off dream but something that might actually be realized.

Could this be achieved even more rapidly? Enter structural genomics. Analysis of the PDB archive showed that, from the point of view of amino acid sequences, there was about a sevenfold sequence redundancy (Hobohm *et al.*, 1992). The number of new protein folds deposited in the PDB archive was relatively low (10% since 2000). At the same time, bioinformatics analysis of the newly determined gene sequences showed that the coverage of all protein families was uneven and incomplete. Would it be possible to very carefully select target sequences that would be novel (<30% sequence similarity), determine their structures and, through homology modeling, obtain structures for the rest of the protein families? Several projects ensued which focused on high-throughput methods for structure determination of these unique targets (*Nature Structure Biology*, 2000; Levitt, 2007). To date, structural genomics centres worldwide have been responsible for adding more than 5500 new structures of which 50.5% are novel.

4. Challenges to the PDB

When the PDB began there were relatively few structures, the structures were small, and only X-ray crystallographic methods were used to determine the structures. Today, the PDB has structures with molecular weights of more than two million, and new structure-determination methods are waiting in the wings to join NMR and cryoEM in the PDB Exchange Dictionary. To meet the challenges of handling these data and making them accessible to a broad community of scientists, methods had to be developed for data representation, acquisition and processing, management and distribution. The information in the PDB must be tied to other data resources, such as GenBank (Benson *et al.*, 2000), UniProt (Apweiler *et al.*, 2004), model organism databases and many, many others. In addition, the global nature of science makes it essential to coordinate these data efforts at an international level.

4.1. Data representation

In addition to the *xyz* coordinates, an entry in the PDB contains information about the chemistry of the macromolecule, the small-molecule ligands, some details of the data collection and structure refinement, and some structural descriptors. In all, a typical PDB entry has about 400 unique items of data. The PDB file format that was devised in 1976 is simple, easy to read by humans and used by many computer applications. However, as the PDB format is based on the original 80 column punched-card format, the number of atoms and number of residues that can be represented is limited. This means that large macromolecular complexes must be represented in more than one data file in the PDB file format. Additionally, there are implicit assumptions made in this file format that limit its use by modern computer applications. In an effort to remedy these shortcomings, the macromolecular crystallographic information file (mmCIF) was created. The mmCIF dictionary contains well over 3000 definitions of every aspect of the crystallographic experiment and results

Figure 2
A selection of the more than 250 icosahedral virus structures currently available in the PDB archive. The representation of these entries has recently undergone a major facelift.

(Fitzgerald *et al.*, 2005). In this format, each data item is completely defined, along with the relationships among the data items. mmCIF is completely computer readable and can be used to create a relational database. To accommodate data from NMR and cryoEM experiments, a PDB Exchange Dictionary (PDBx) was created that has the same syntax as mmCIF but contains all the definitions needed to handle the data that are now part of the PDB (Westbrook, Henrick *et al.*, 2005). In addition, the mmCIF data files have been translated into PDBML-XML files that can be managed by off-the-shelf tools (Westbrook, Ito *et al.*, 2005).

4.2. Data acquisition and processing

In the earliest days of the PDB at BNL, data were sent by mail on either punched cards or magnetic tape. A form was completed that contained information about the structure. Annotators examined the data and obvious errors were corrected. Checking the geometry was done with a set of special-purpose computer programs. Letters were sent through the post office to authors with reports of the checks and, after author agreement, the data were entered in the computer archive. It was not until the 1990's that fully electronic submission was possible. *AutoDep* allowed the author to submit files online (Lin *et al.*, 2000). Some checks were done automatically and others were done manually.

When the PDB changed management in 1998, data acquisition and processing began to evolve. EBI reengineered *AutoDep* and made it more automatic (Keller *et al.*, 1998). The RCSB PDB created the *AutoDep* Input Tool (ADIT) (Berman *et al.*, 2000) which was based on mmCIF. PDBj also adopted this tool for data collection. The newest wwPDB member, the BMRB (http://www.bmrb.wisc.edu), uses ADIT-NMR to collect experimental and coordinate data at the same time. Data processing, although done with different tools around the world, uses the same principles and algorithms. All

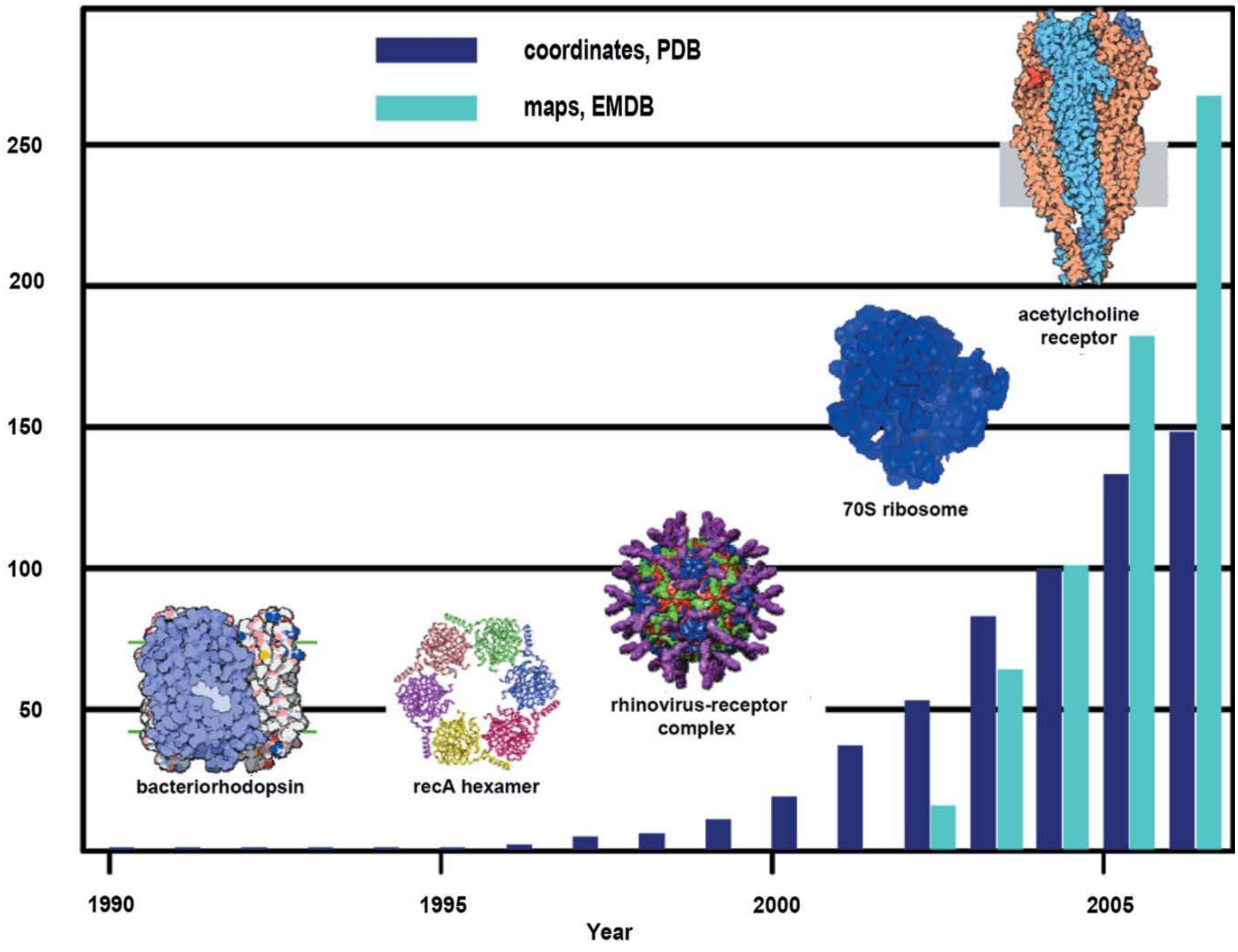

Figure 3
Growth of the number of cryoEM structures in the PDB and the number of related maps in the Electron Microscopy Database (EMDB) (Henrick *et al.*, 2003).

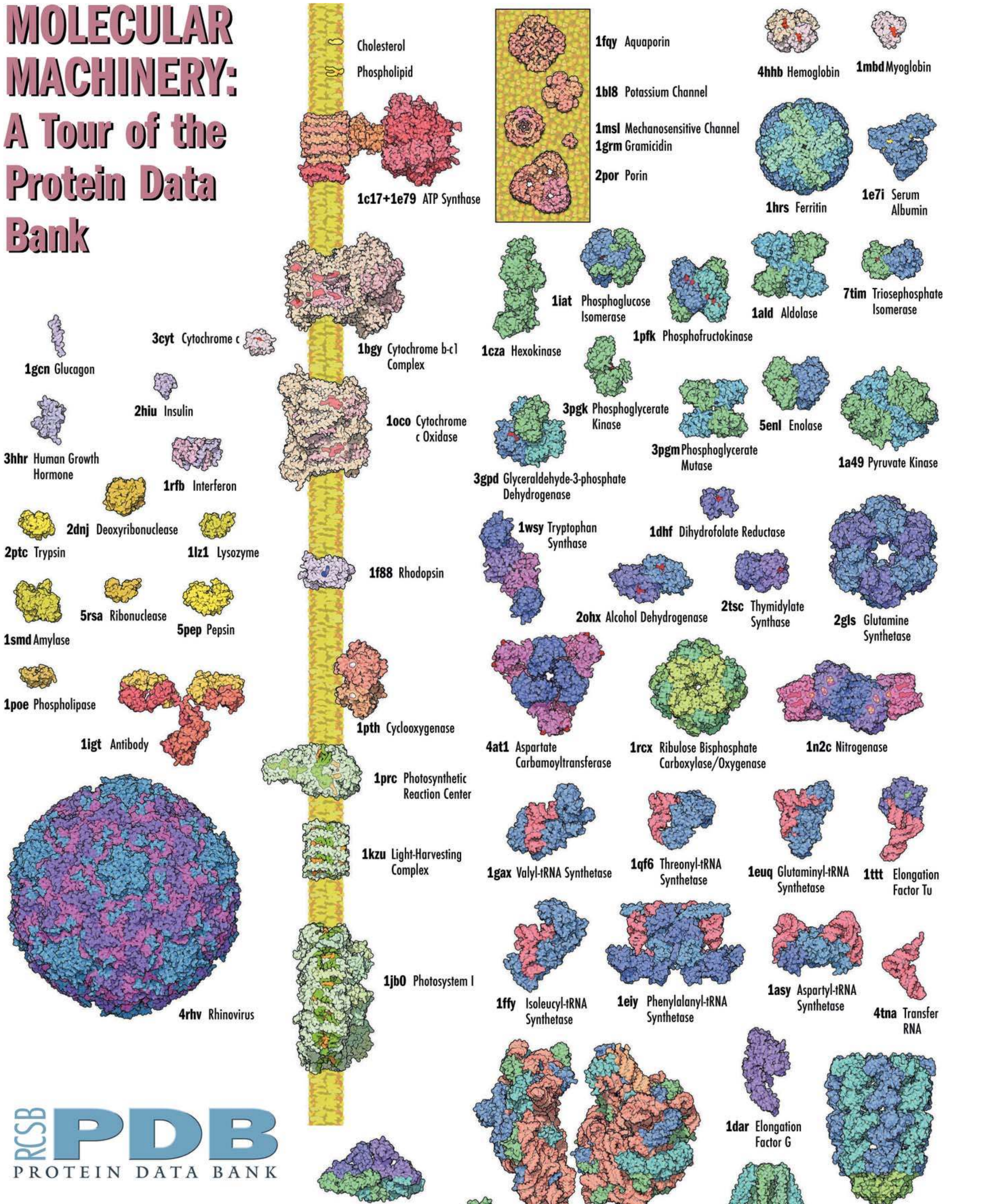
MOLECULAR MACHINERY: A Tour of the Protein Data Bank
Cholesterol
Phospholipid
1c17+1e79 ATP Synthase
1fqy Aquaporin
1bl8 Potassium Channel
1msl Mechanosensitive Channel
1grm Gramicidin
2por Porin
4hhb Hemoglobin
1mbd Myoglobin
1hrs Ferritin
1e7i Serum Albumin
1bgy Cytochrome b-c1 Complex
1cza Hexokinase
1iat Phosphoglucose Isomerase
1pfk Phosphofructokinase
1ald Aldolase
7tim Triosephosphate Isomerase
3cyt Cytochrome c
1gcn Glucagon
2hiu Insulin
3hhr Human Growth Hormone
1rfb Interferon
1oco Cytochrome c Oxidase
3pgk Phosphoglycerate Kinase
3pgm Phosphoglycerate Mutase
5enl Enolase
1a49 Pyruvate Kinase
3gpd Glyceraldehyde-3-phosphate Dehydrogenase
2dnj Deoxyribonuclease
2ptc Trypsin
1lz1 Lysozyme
1f88 Rhodopsin
1wsy Tryptophan Synthase
1dhf Dihydrofolate Reductase
2ohx Alcohol Dehydrogenase
2tsc Thymidylate Synthase
2gls Glutamine Synthetase
5rsa Ribonuclease
1smd Amylase
5pep Pepsin
1poe Phospholipase
1igt Antibody
1pth Cyclooxygenase
4at1 Aspartate Carbamoyltransferase
1rcx Ribulose Bisphosphate Carboxylase/Oxygenase
1n2c Nitrogenase
1prc Photosynthetic Reaction Center
1kzu Light-Harvesting Complex
1gax Valyl-tRNA Synthetase
1qf6 Threonyl-tRNA Synthetase
1euq Glutaminyl-tRNA Synthetase
1ttt Elongation Factor Tu
4rhv Rhinovirus
1jb0 Photosystem I
1ffy Isoleucyl-tRNA Synthetase
1eiy Phenylalanyl-tRNA Synthetase
1asy Aspartyl-tRNA Synthetase
4tna Transfer RNA
RCSB PDB PROTEIN DATA BANK
1dar Elongation Factor G
1efu Elongation Factor Tu and Ts
2cpl Proline cis/trans Isomerase
1fjf+1jj2 Ribosome
1fxk Prefoldin
1aon Chaperonin GroEL/ES

Figure 4
A look at the diversity of structures in the PDB archive. These images, shown to scale, were created by David S. Goodsell (The Scripps Research Institute), who also writes and illustrates the RCSB PDB's *Molecule of the Month* feature. An expanded version of this figure is available for download from http://www.pdb.org.

PDB files are checked for accuracy of the geometry, chemistry of the polymer and ligands, nomenclature, and the likely biological assembly. In recent years, structure factors and NMR constraint files are deposited with the majority of data files so that now it is possible to calculate the agreement with experimental data using *SFCheck* (Vaguine *et al.*, 1999). To ensure complete uniformity, the wwPDB has reviewed and documented all data processing practices among the member sites. The wwPDB has also taken on the task of reviewing all the files in the archive and has very recently created a new, more uniform, archive at ftp://ftp.wwpdb.org (Henrick *et al.*, 2008)

4.3. Data distribution and query

In the early days of the PDB, data were distributed *via* magnetic tape and later by CD-ROM. Now there is an ftp site that contains the data in three formats: PDB, mmCIF-PDBx and PDBML-XML. Distributing these data on media would require more than 15 DVD disks. The ftp site is updated weekly and each wwPDB center maintains a mirror of the site. The RCSB PDB website alone is accessed by about 100000 unique visitors per month from nearly 140 different countries. More than 500 GB of data are transferred each month. On a typical weekday, two pages from the site are viewed every second. Data are accessed *via* the website, ftp servers (supporting ftp and rsync access), web services and RSS feeds.

Until the 1990's, interactive query was not possible from a central site. The World Wide Web changed all that. The first web-accessible interface was made available at BNL and allowed many useful queries (Prilusky, 1996). Now the wwPDB sites offer numerous services including simple and complex searches as well as a variety of visualization methods. In addition, the RCSB PDB (http://www.pdb.org) provides browsing capabilities across external resources, a Structure Summary page for each entry and a *Molecule of the Month* feature highlighting a particular structure (Berman *et al.*, 2000). PDBj (http://www.pdbj.org) provides several services, including Alignment of Structural Homologs (Standley *et al.*, 2007). The services of the MSD-EBI (http://www.ebi.ac.uk/msd) include analyses of macromolecule ligand interactions, statistical analyses and residue-based analyses (Golovin *et al.*, 2004).

The accessibility of the data and the growing importance of understanding the data has meant that the PDB's user community has grown from the community of crystallographers that banded together to form the archive. The PDB archive is a critical resource for researchers in academia and industry, working in subjects such as structural biology, computational biology, biophysics, biochemistry, genetics, cell biology and molecular biology. The PDB is also a tool used by educators and students from middle school to graduate school.

5. Applications of PDB data in academia and industry

The PDB is very widely used. For example, an average of 211515 files are downloaded from the ftp site each day. To date, there are more than 5000 references to the first RCSB PDB publication (Berman *et al.*, 2000), making it one of the most-cited papers in all of biology. While the early users of the PDB were mostly crystallographers who used the resource to store their data and to review other structures for comparison with their own, now more than half of the papers that cite this one publication alone describe bioinformatic and computational analyses of structural data. Enormous efforts are under way to be able to understand protein folding so that perhaps someday it will be possible to predict structure from sequence (Moult, 2005). While the PDB is considered an archival data resource that stores and distributes primary data, there are hundreds of derivative databases that catalog the data in different ways. For example, CATH (Orengo *et al.*, 1997) and SCOP (Conte *et al.*, 2000) provide classifications of the folds found in proteins. More recently, there have also been efforts to understand protein–protein interactions (Janin *et al.*, 2003) and again to try to predict these. There are specialty databases such as the Nucleic Acid Database (Berman *et al.*, 1992) and the HIV Protease Structural Database (Ravichandran *et al.*, 2002) that create in-depth resources for researchers in nucleic acids and HIV, respectively.

When the PDB updates the ftp site each week, most pharmaceutical companies download the new data for inclusion in their own in-house databases. These structural data are used to aid the discovery of new pharmaceuticals. Indeed, the ready availability of the structure of HIV protease (Navia *et al.*, 1989) enabled many companies to concentrate their efforts on the development of effective protease inhibitors that are now the basis of AIDS treatment.

6. Future of the PDB

The PDB archive is a key example of a community resource that has evolved over its 36 year history. Its evolution has been driven by changes in science and technology used to determine structures, the nature of the structures that are determined, community attitudes about data sharing, and the nature of the communities that are interested in structural data.

In the short term, there will be several new challenges. There needs to be better representation for disordered structures, for X-ray structures refined with multiple models (Furnham *et al.*, 2006) and for very large macromolecular complexes. New annotations will be required to describe function. All of these changes must be done in consultation with depositors, software developers and users of the data. Annotation practices used by the different data centers will continue to be examined and standardized so as to keep the archive uniform. It is likely that the wwPDB sites will develop a single deposition system. In the long term, the wwPDB will be able to develop new joint services for analysis and browsing of the rich data contained within the archive.

As the PDB continues to evolve, in addition to being able to use these data to perhaps predict structure, an even greater challenge will be to determine function by knowing the structure. Once accomplished, the long-term vision of enabling a molecular view of biology and medicine will become a reality.

Many people have been key in the development and maintenance of the PDB in its 36 year history. These include previous directors Tom Koetzle and Joel L. Sussman, and the BNL staff including Enrique Abola and Frances Bernstein. John Westbrook and Phil Bourne have been instrumental to the development of the RCSB PDB resources. MSD-EBI leader Kim Henrick and PDBj leader Haruki Nakamura have worked hard to ensure that the PDB remains a global resource. The many staff members of the wwPDB data centers continue to contribute to the continued viability and quality of the PDB. The RCSB PDB is supported by DBI 0312718.

References

Allen, F. H., Kennard, O., Motherwell, W. D. S., Town, W. G. & Watson, D. G. (1973). *J. Chem. Doc.* **13**, 119–123.

Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N. & Yeh, L. S. (2004). *Nucleic Acids Res*. **32**, Database issue, D115–D119.

Arnold, E. & Rossmann, M. G. (1988). *Acta Cryst.* A**44**, 270–283.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., Rapp, B. A. & Wheeler, D. L. (2000). *Nucleic Acids Res.* **28**, 15–18.

Berman, H. M., Henrick, K. & Nakamura, H. (2003). *Nature Struct. Biol.* **10**, 980.

Berman, H. M., Olson, W. K., Beveridge, D. L., Westbrook, J., Gelbin, A., Demeny, T., Hsieh, S. H., Srinivasan, A. R. & Schneider, B. (1992). *Biophys. J.* **63**, 751–759.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F. Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.

Brünger, A. T. (1990). *X-PLOR. A System for Crystallography and NMR*, Version 2.1. New Haven, CT, USA: Yale University.

Cold Spring Laboratory Press (1972). Cold Spring Harbor Symposia on Quantitative Biology, Vol. 36.

Conte, L., Bart, A., Hubbard, T., Brenner, S., Murzin, A. & Chothia, C. (2000). *Nucleic Acids Res.* **28**, 257–259.

Fitzgerald, P. M. D., Westbrook, J. D., Bourne, P. E., McMahon, B., Watenpaugh, K. D. & Berman, H. M. (2005). *International Tables for Crystallography*, Vol. G. *Definition and Exchange of Crystallographic Data*, edited by S. R. Hall & B. McMahon, ch. 4.5, *Macromolecular Dictionary (mmCIF)*, pp. 295–443. Dordrecht: Springer.

Furnham, N., Blundell, T. L., DePristo, M. A. & Terwilliger, T. C. (2006). *Nature Struct. Mol. Biol.* **13**, 184–185; discussion p. 185.

Golovin, A., Oldfield, T. J., Tate, J. G., Velankar, S., Barton, G. J., Boutselakis, H., Dimitropoulos, D., Fillon, J., Hussain, A., Ionides, J. M., John, M., Keller, P. A., Krissinel, E., McNeil, P., Naim, A. *et al.* (2004). *Nucleic Acids Res*. **32**, Database issue, D211–D216.

Helliwell, J. (1983). *Acta Radiol. Suppl.* **365**, 35–37.

Hendrickson, W. A. (1991). *Science*, **254**, 51–58.

Hendrickson, W. A. & Konnert, J. H. (1981). *PROLSQ*, Vol. 1, B*iomolecular Structure, Conformation, Function and Evolution*, edited by R. Srinivasan, E. Subramanian & N. Yathindra, pp. 43–57. Oxford: Pergamon Press.

Henrick, K., Feng, Z., Bluhm, W., Dimitropoulos, D., Doreleijers, J. F., Dutta, S., Flippen-Anderson, J. L., Ionides, J., Kamada, C., Krissinel, E., Lawson, C. L., Markley, J. L., Nakamura, H., Newman, R., Shimizu, Y. *et al.* (2008). *Nucleic Acids Res.* Database Issue. In the press.

Henrick, K., Newman, R., Tagari, M. & Chagoyen, M. (2003). *J. Struct. Biol.* **144**, 228–237.

Hobohm, U., Scharf, M., Schneider, R. & Sander, C. (1992). *Protein Sci.* **1**, 409–417.

International Union of Crystallography (1989). *Acta Cryst.* A**45**, 658.

Jancarik, J. & Kim, S.-H. (1991). *J. Appl. Cryst.* **24**, 409–411.

Janin, J., Henrick, K., Moult, J., Eyck, L. T., Sternberg, M. J., Vajda, S., Vakser, I. & Wodak, S. J. (2003). *Proteins*, **52**, 2–9.

Jones, T. A. (1978). *J. Appl. Cryst.* **11**, 268–272.

Jones, T. A., Zou, J.-Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Acta Cryst*. A**47**, 110–119.

Keller, P. A., Henrick, K., McNeil, P., Moodie, S. & Barton, G. J. (1998). *Acta Cryst.* D**54**, 1105–1108.

Kennard, O., Watson, D. G. & Town, W. G. (1972). *J. Chem. Doc.* **12**, 14–19.

Lehmann, M. S., Koetzle, T. F. & Hamilton, W. C. (1972). *J. Am. Chem. Soc.* **94**, 2657–2660.

Levitt, M. (2007). *Proc. Natl Acad. Sci. USA*, **104**, 3183–3188.

Lin, D., Manning, N. O., Jiang, J., Abola, E. E., Stampf, D., Prilusky, J. & Sussman, J. L. (2000). *Acta Cryst.* D**56**, 828–841.

Meyer, E. F. (1997). *Protein Sci.* **6**, 1591–1597.

Moore, P. (2001). *Biochemistry*, **40**, 3243–3250.

Moult, J. (2005). *Curr. Opin. Struct. Biol.* **15**, 285–289.

Nature Structural Biology (2000). *Archive – Nature Structural Biology*, 7:11s, http://www.nature.com/nsmb/journal/v7/n11s/index.html.

Navia, M. A., Fitzgerald, P. M., McKeever, B. M., Leu, C. T., Heimbach, J. C., Herber, W. K., Sigal, I. S., Darke, P. L. & Springer, J. P. (1989). *Nature (London)*, **337**, 615–620.

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. (1997). *Structure*, **5**, 1093–1108.

Prilusky, J. (1996). *OCA, a Browser-Database for Protein Structure/Function*, http://bip.weizmann.ac.il/oca and mirrors worldwide.

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

Ravichandran, V., Vondrasek, J., Gilliland, G., Bhat, T. N. & Wlodawer, A. (2002). CSB Proceedings of the IEEE Computer Society Conference on Bioinformatics 340. Washington: IEEE Computer Society.

Standley, D. M., Toh, H. & Nakamura, H. (2007). *BMC Bioinformatics*, **8**, 116.

Vaguine, A. A., Richelle, J. & Wodak, S. J. (1999). *Acta Cryst.* D**55**, 191–205.

Westbrook, J., Henrick, K., Ulrich, E. L. & Berman, H. M. (2005). *International Tables for Crystallography*, Vol. G. *Definition and Exchange of Crystallographic Data*, edited by S. R. Hall & B. McMahon, ch. 3.6.2, *The Protein Data Bank Exchange Data Dictionary*, pp. 195–198. Dordrecht: Springer.

Westbrook, J., Ito, N., Nakamura, H., Henrick, K. & Berman, H. M. (2005). *Bioinformatics*, **21**, 988–992.