RCSB Protein Data Bank Advisory Committee
Report of November 1, 2016 Annual Meeting
National Science Foundation, Arlington, VA

Chair: Cynthia Wolberger

Membership: Paul Adams, R. Andrew Byrd, Bridget Carragher, Wah Chiu, Kirk Clark, Paul Craig, Roland L. Dunbrack, Jr., Thomas E. Ferrin, Catherine E. Peishoff, Sue Rhee, Andrej Sali (absent), Torsten Schwede, Jill Trewhella and Cynthia Wolberger

RCSB PDB AC E-mail Addresses:
cwolberg@jhmi.edu, PDAdams@LBL.gov, bcarr@nysbc.org, byrda@mail.nih.gov, wah@bcm.edu, kirk.clark@novartis.com, paul.craig@rit.edu, roland.dunbrack@fccc.edu, tef@cgl.ucsf.edu, catherine.E.Peishoff@gsk.com, srhee@carnegiescience.edu, sali@salilab.org, torsten.schwede@unibas.ch, j.trewhella@mmb.usyd.edu.au

US Government Representatives: Wilson A. Francisco (NSF), Ranajeet Ghose (NSF), Lin He (NSF), Roland F. Hirsch (DOE), J. Randy Knowlton (NCI, via phone), Catherine Lewis (NIH, via phone), Ramana Madupu (DOE), Peter McCartney (NSF), David Alexander Rockcliffe (NSF), Engin Serpersu (NSF), Ward W. Smith (NIH), Amy Swain (DOE), Jennifer Weller (NSF)

US Government Agency Representative E-mail Addresses:
pmccartney@nsf.org, wfrancis@nsf.gov, RGHOSE@nsf.gov, lhe@nsf.gov, Roland.Hirsch@science.doe.gov, Jk339o@nih.gov, lewisc@nigms.nih.gov, ramana.madupu@science.doe.gov, pmccartn@nsf.gov, drockcli@nsf.gov, ESERPERS@nsf.gov, smithwar@nigms.nih.gov, Amy.Swain@science.doe.gov, jweller@nsf.gov

RCSB Leadership: Stephen Burley, Helen Berman

RCSB Leadership E-mail Addresses:
sburley@proteomics.rutgers.edu, berman@rcsb.rutgers.edu

Executive Summary

The Advisory Committee to the Research Collaboratory for Structural Bioinformatics (RCSB) - met in Arlington, VA on 1st November 2016 to consider management and enhancement of the Protein Data Bank (PDB).

Agenda items included
(1) Overview/State of the PDB;
(2) Data In: OneDep, Data Standards, Infrastructure, Plans Forward;
(3) Data Out: Access, Exploration, and Metrics
(4) Outreach;
(5) Education;
(6) Funding and Sustainability
(7) 2015 Advisory Committee Response;
(8) Matters Arising; and
(9) General Discussion

The purpose in holding the meeting at the National Science Foundation (NSF) was to facilitate participation by representatives of the NSF, National Institutes of Health (NIH) and the Department of Energy (DOE), who jointly fund the RCSB PDB. The Committee and the RCSB leadership were very grateful that multiple representatives from all three agencies accepted the invitation to participate, and appreciate their many thoughtful contributions to the discussion.

The meeting was opened by Dr. Stephen Burley, who gave an overview of the past year's activities and current state of the RCSB PDB. Drs. Jasmine Young, John Westbrook, Peter Rose, Andreas Prlić, Helen Berman and Shuchi Datta presented summaries of recent RCSB activities. The meeting closed with a presentation by Dr. Burley of the RCSB PDB response to the recommendations of the 2015 RCSB PDB AC, as well as a continued discussion of funding and sustainability.

A critically important accomplishment of the past year was the successful implementation of the OneDep deposition system across all three wwPDB sites (RCSB, PDBe, and PDBj). This collaborative endeavor shared by all three sites had been subjected to significant delays, thus endangering the ability of the RCSB PDB and partner sites to cope with ever-increasing numbers and complexities of deposition. The Committee commends the entire RCSB PDB team for their contributions and leadership in resolving the issues causing the delays. We particularly note the superlative job done by Dr. Jasmine Young as the project manager. The Committee also congratulates the RCSB PDB for implementing numerous additional improvements over the past year, including rollout of the PDBx/mmCIF format to accommodate larger structures and more detailed metadata, the Ligand Explorer tool, the new NGL web viewer, and options to color-code structures by their fit to electron density. The Committee encourages development of analogous methods for visualizing NMR structure quality.

The RCSB PDB continues to do an outstanding job in outreach and education. The education plan, which is focused on developing course curricula and materials for teachers, represents an excellent approach to maximizing the impact of the education program. The increasing popularity of the high school video challenge is another testament to the success of RCSB outreach. The Committee encourages the RCSB to include in its popular Molecule of the Month feature structures that are relevant to the research missions of the DOE and NSF, along with its already highly regarded offerings that are focused on human health and disease. The Committee was highly enthusiastic about the ongoing effort, spearheaded by Dr. Helen Berman, to create a documentary on the molecular basis of HIV in collaboration with documentary filmmakers at the University of Southern California School of Cinematic Arts.

Obtaining long-term, stable funding is crucial for enabling the RCSB PDB to serve the needs of the scientific, medical, industry and education communities. The Advisory Committee is grateful for the ongoing commitment of the NIH, NSF and DOE, which was reinforced by their participation in this year's meeting. As planning for the 2018 grant renewal proceeds, the Committee encourages the RCSB leadership to gather comprehensive metrics as well as specific examples of the profound impact of the PDB on research funded by all three agencies. In the case of the DOE there are ample opportunities that can be highlighted given the agency's focus on environmental molecular science and the fact that it operates much of the major infrastructure used by the structural biology community. The NSF's focus on the frontiers of biological knowledge and on increasing our understanding of complex systems is ripe for outreach activities and is also highly relevant to the integrative/hybrid structure initiative.

Responses to 2015 RCSB PDB AC Recommendations

PDBAC: Complete OneDep tool in collaboration with other wwPDB sites
Response: Accomplished and successfully launched.

PDBAC: Investigate reasons users replace coordinates after deposition
Response: Implemented pre-deposition server to allow users to check coordinates and data prior to full deposition.

PDBAC: Remediate carbohydrates, residual B factors and crystal orientation.
Response: Remediation priority for 2016 shifted to translating EM data files to the current PDBx/mmCIF dictionary in order to support the OneDep system. This process is nearly complete, to be released by end of 2016.

PDBAC: Obtain additional funding for handling structures determined by integrative/hybrid methods.
Response: Have had a number of successes to date and a number of new submissions have been made that are under review. Additional funding is essential to fully implement this suggestion.

PDBAC: Initiate pilot project to explore disease-causing mutations and measure interest in research and educational communities before broadening project's scope.
Response: Completed pilot project on EGF receptor.

PDB Metrics

In aggregate, 10,958 depositions were deposited and processed between January 1st and December 31st 2015 with a two-week average turnaround, an increase from the 10,364 entries deposited in 2014. Based upon the number of entries deposited in 2016 to date, it is estimated that 11000 entries will be deposited this year.

Breakdown of depositions by discipline in calendar 2015 was as follows:

X-ray: 10,169 (93% of entries deposited, up from 9586 in 2014)
NMR: 510 (5%, down from 515 in 2014)
EM: 255 (2%, up from 240 in 2014)
Other: 24 (0.2%, up from 23 in 2014)

Breakdown of depositions by wwPDB processing site in calendar 2015 was as follows:

RCSB PDB: 4845 (44%)
PDBj: 2100 (19%)
PDBe-EBI: 4013 (37%)

Breakdown of depositors by location in calendar 2014 was as follows:

North America 39%
Europe 34%
Asia 20%

Industry	2%
South America	1%
Australasia	4%
Africa	<1%

During 2015, RCSB PDB's website at http://rcsb.org was visited each month by an average of ~316,000 unique visitors and ~741,000 unique visits. A total of 35.260 TB of data were accessed.

During the same period, more than 534 million data files were downloaded from the PDB archive *via* the wwPDB member FTP and websites (RCSB PDB: 367,149,527; PDBe: 89,671,549; PDBj: 77,518,795).

2016 RCSB PDB AC Discussion

Overview of PDB

Dr. Stephen Burley presented an overview of RCSB PDB activities over the past year as well as the challenges that lie ahead. The number of depositions has continued to grow at the rate of about 10% per year, with the total number exceeding 124,000 as of the day of the AC meeting. The important role that the RCSB PDB plays as steward of the archive, and in ensuring data security and disaster recovery, is critical to global progress in biology, biotechnology development, and biomedical applications including drug discovery. As a compelling illustration, replicating the contents of the PDB archive would cost $12 billion, assuming a conservative estimate of $100,000 per structure. Dr. Burley also noted that the PDB has, from its inception, adhered to the recently articulated FAIR guiding principles for scientific data management and stewardship: Findability, Accessibility, Interoperability and Reusability. The Committee considers this as yet another illustration of forward-thinking data management throughout the history of PDB leadership.

The continued growth of structures determined by hybrid methods presents both a challenge and an opportunity for the PDB, as it must adapt data-in, data validation, archiving and data-out to deal with structures that are typically large, complex and determined with multiple kinds of experimental techniques. This will require significant additional external funding, which the Committee very much hopes can soon be obtained in order to meet these important goals.

Data In: OneDep, Data Standards, Infrastructure, Plans Forward

Dr. Jasmine Young updated the Committee on the outstanding progress over the past year in completing and rolling out the automated deposition and annotation system, formerly denoted D&A, which has been renamed OneDep. The Committee salutes Dr. Young's leadership, who was appointed project manager one year ago. Dr. Young played a crucial role in coordinating the efforts of an international team of scientists, programmers and software developers and bringing the project to a successful conclusion. This hugely important achievement will allow the wwPDB partners to balance the workload more effectively. Other advantages of OneDep is that it captures data more completely, standardizes files, improves efficiency and consistency, allows file replacement, and has standalone validation for all methods at various stages of deposition. One interesting consequence of the validation feedback is the striking rate of coordinate replacement, with ~ 25% of depositors replacing coordinates in response to validation feedback during processing. While this ultimately leads to higher quality structures and data consistency, it also increases the burden on curators. The Committee hopes that the

release of a pre-deposition server will encourage depositors to check coordinates in advance of depositing them. In addition, the Committee hopes that more journals will embrace policies requiring validation reports at the time of manuscript submission, as has been done at Journals including the *Nature* family, *Acta Crystallographica D* and *F, J. Biological Chemistry, J. Immunology, eLIFE, Structure,* and *Angewandte Chem. Int. Ed. Engl.* Overall, the positive impact of validation has been clear from the improvement in geometry and clash scores in structures deposited since the advent of the wwPDB validation report. The plans for carbohydrate remediation will further improve standardization, and thus facilitate successful data searches.

Dr. John Westbrook gave an overview of the implementation of the PDBx/mmCIF coordinate format, which can accommodate large structures and more information than the legacy PDB format. This has facilitated efforts on ligand validation, as outlined in the recent report in *Structure* covering the 2015 ligand validation workshop. One important development has been the implementation of the Ligand Explorer option to display electron density for ligands. The Committee endorses continued plans to validate and standardize bound ligands.

The Committee was updated about progress in developing methods to model structures derived from hybrid methods, a pilot project funded by an NSF EAGER grant. Continued progress and increased grant support will be critical given the increasing number of structures determined through integrative/hybrid methods.

Data Out: Access, Exploration and Metrics

Dr. Peter Rose provided an overview of the website and how Google analytics are used to track patterns of usage. Dr. Rose introduced some of the most recently developed website features, including the ability to search PDB-101 content and improvements to the search results page. The Committee was gratified to see that last year's overhaul of the RCSB PDB website, which was launched in October 2015, has resulted in significantly increased usage of features such as structure similarity, sequence similarity and experimental information. These statistics demonstrate the importance of the significant effort that went into making the website easier to navigate and view. Several highly important and useful new features involve integrating proteomic, genomic and experimental data with structural information, as well as the pathway view feature. The new NGL web viewer and MMTF data format are outstanding additions that greatly enhance visualization and download speed, thanks to use of compressed binary data. The Committee was also impressed with other new features that display ligand electron density and structures color-coded by real space fit to electron density, and supports ongoing efforts to create comparable tools for structures determined by NMR.

Dr. Andreas Prlić provided measures of website impact that were obtained using web analytics. The RCSB PDB website has a remarkable 350,000 unique monthly users, with over 1 million unique users annually, a testament to the tremendous impact of RCSB resources. Interestingly, access by FTP is double that of webpage access, indicating that a majority of the coordinate data is being utilized by computer programs that directly access the RCSB data. The Committee encourages the RCSB team to continue to think of ways to make users aware of new features in order to maximize their impact. Looking ahead, the Committee endorses the plans for future improvements that address the needs of both expert and non-expert communities, as well as hardware improvements that can accommodate continued growth of user needs as well as new features.

Outreach: Using film to connect structural biology with real-world problems

Dr. Helen Berman treated the committee to a preview of a documentary whose goal is to explain to non-structural biologists how insights obtained from structural biology can be used to understand and treat real-world problems. Dr. Berman spent a sabbatical at the University of Southern California (USC) School of Cinematic Arts, where she initiated a collaboration with documentary filmmakers on a film about HIV, which combines an emotionally compelling narrative with molecular animations and expert commentary. The Committee gave the preview a unanimous “thumbs up.” The Committee was enormously impressed by the exciting and ingenious new ways HIV was explained from a combined human, medical and structural perspective, and looks forward to seeing the completed film. The Committee salutes Dr. Berman for her creativity and resourcefulness in forging an alliance with experienced filmmakers to develop more effective ways to educate the public about human disease and its treatment, while at the same time conveying the underlying science in a way that can be understood by the public.

Education

Dr. Shuchismita Dutta provided an update on the education and outreach plan of the RCSB PDB. The Committee was impressed by the development of course materials organized around specific themes and directed towards either undergraduates or high school students. The choice to develop curricula and other materials centered on diabetes is an excellent one that is likely to engage students on a personal level given the large affected population. The well thought-out plans to test and evaluate the effectiveness of educational materials is a strength of the approach that will ensure maximal impact of these efforts. The more than two-fold increase in the number of high school students participating in this year’s video challenge is a testament to the growing impact of this creative way of engaging students in activities that utilize structural information. The redesign of the popular PDB-101 portion of the website was viewed as outstanding and will make this feature even more accessible and useful to students and the broader non-expert community.

Funding and Sustainability

Dr. Stephen Burley led the final discussion of current and future support for the RCSB PDB, including plans for submission of the grant renewal in 2018. The Committee feels that the RCSB team’s ability to implement multiple improvements while serving an ever-growing user base is remarkable despite a 14.8% decrease in purchasing power over the last 12 years, and a 3.5% budget cut in real dollars since 2013. While the RCSB PDB has been able to realize some cost savings due to efficiency, most notably through the joint development of automated deposition, sustained funding is absolutely critical. Another issue confronting the RCSB PDB is the lack of a funding mechanism for the wwPDB, the consortium that coordinates efforts among the three current partners, RCSB PDB, PDBe, PDBj and BMRB. With the planned expansion of the wwPDB to include partners in China and India, it will be essential to obtain funding for the operation of what is an increasingly complex international organization involving independently operated and governed entities that must work together in a common framework with common purpose.

The Committee hopes that the wwPDB Foundation, which was established as a 501(c)3 organization to fund the outreach efforts of the wwPDB, will be successful in meeting this need.

The Committee is deeply concerned about the sustainability of funding for the RCSB PDB and fears that the central importance of this archive to fundamental and translational research is not

sufficiently appreciated by funders. Primary databases like the PDB and GenBank are absolutely essential to a broad swath of researchers. Suggestions that the RCSB PDB seek alternative models of funding, such as subscriptions, not only ignores the broad use of the PDB by non-specialists but also would be an explicit violation of the agreement that unites the wwPDB partners. The Committee embraced the idea that NIH consider funding the PDB from the office of the Director, rather from NIGMS or another institute, in order to remove funding decisions from direct competition with research grants, and to provide a more stable funding stream. The Committee encourages the PDB leadership to assemble comprehensive metrics that show the utility and impact of the PDB on research funded by the NSF, NIH and DOE. To aid in this process, the RCSB PDB should consider new methods for capturing information on the impact of the PDB. One suggestion was to offer incentives for users to register for an account and to make use of the user community to gather information about the impact of the PDB and advertise new developments and features. The RCSB PDB could also consider partnering with scientific societies and organizations such as ASBMB and Howard Hughes Medical Institute to gather information on the impact of PDB data and tools outside the structural biology community. In addition, the Committee suggests that the RCSB PDB utilize its PDB-101 and Molecule of the Month feature to highlight advances relevant to research funded by the NSF and DOE, in addition to the many outstanding examples of medical importance already covered by the educational portion of the website.

In light of the current highly competitive funding landscape, the Committee suggests that the RCSB leadership establish priorities with respect to applications for additional funding. While renewing the core funding is of the highest importance, some additional grants are needed to underwrite specific projects. The Committee encourages the RCSB leadership to weigh the substantial effort needed for applying for each additional grant in order to focus only on applications that have the greatest potential impact.