RCSB Protein Data Bank Advisory Committee
Report of November 3, 2015 Annual Meeting
Rutgers University, New Brunswick, New Jersey

Chair: Cynthia Wolberger

Membership: Paul Adams, R. Andrew Byrd, Wah Chiu (absent), 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

US Government Representatives: Peter McCartney (NSF representative, present for Skype discussion)

RCSB Leadership: Stephen Burley, Helen Berman

RCSB PDB AC E-mail Addresses:
cwolberg@jhmi.edu, PDAdams@LBL.gov, 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 Agency Representative E-mail Addresses:
pmccartney@nsf.org

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 New Brunswick, New Jersey on 3rd November 2015 to consider management and enhancement of the Protein Data Bank (PDB).

Agenda items included
(1) Responses to 2014 RCSB PDB AC Recommendations;
(2) State of the PDB;
(3) Update on Integrative and Hybrid Methods
(4) Data In: Deposition and Annotation;
(5) Data Out: Access and Exploration;
(6) Education plan
(7) The PDB-101 website;
(8) Management issues;
(9) Discussion with Dr. Peter McCartney, NSF; and
(10) Matters arising.

The meeting was held in the Rutgers University Center for Integrative Proteomics and opened by Dr. Stephen Burley, who gave an overview of the past year's activities and current state of the RCSB PDB. Dr. Burley welcomed the new members of the Advisory Committee and outlined the new policy of appointing Members to 3-year renewable terms. Burley outlined the responses to the 2014 RCSB PDB AC Recommendations and updated the committee on progress towards

completing the next version of the Deposition and Annotation (D&A) tool in partnership with PDBe. A summary of recent activities was subsequently provided by Berman, Young, Westbrook, Rose, Prlić, Dutta and Goodsell.

The Committee felt that the RCSB PDB has done a superlative job in addressing the issues raised in the 2014 PDB AC report. The Committee praises the leadership of Drs. Burley and Berman, whose well-managed team is effectively meeting new challenges and has been highly successful for obtaining additional funding for targeted initiatives, as recommended. The Committee was very enthusiastic about the restructuring of the Outreach and Education plan, one of the recommendations in last year's report. The new Education plan is focused and leverages successful components of previous education efforts to achieve maximal impact. The Committee encourages the RCSB PDB to continue to monitor the impact and effectiveness of the new education plan.

Together with impressive gains in efficiencies thanks to the automated D&A tool, the RCSB PDB is in an excellent position to deal with increasing numbers of depositions and to meet the challenges of handling more complex depositions of structures determined by hybrid methods. Accompanying the dramatic reduction in turnaround for coordinate deposition is a marked increase in coordinate replacement, which could be an indication of improvement in the quality of the model in light of validation information provided during the deposition process. The Committee recommends investigating the reasons that users replace coordinates after the initial deposition and identify mechanisms that would encourage researchers to validate coordinates and data prior to beginning the deposition. As part of this effort, the Committee recommends making the validation pipeline software through a web-accessible interface as well as a standalone downloadable version. The Committee emphasizes the **critical importance of completing version 2.0 of the D&A tool, which will be essential to meeting future demands across all four wwPDB sites**. It is thus a matter of deep concern to the Committee that completion of D&A 2.0 has been delayed by over a year. The Committee very much hopes that the new management agreement and new deadlines agreed upon by the wwPDB collaboration will result in release of D&A 2.0 in early 2016.

The Committee endorses several proposals by RCSB PDB to improve the quality of deposited data and enhance the ability of users to connect structural data to information on biological function. These include plans to remediate carbohydrates, residual B factors and crystal orientation, as previously discussed, as well as a proposal to include visualization of ligand electron density. The Committee also supports the proposal to map structures to protein families and to biological pathways, which will be highly useful to the general user community. It might be useful to assess which resources have the most intuitive representation of biological pathways for the bulk of PDB users.

The Committee emphasizes once again the importance of securing stable, long-term funding for RCSB PDB to serve the needs of the scientific, medical, industrial and education communities. The Committee is grateful to the NIH, NSF and DOE for their long-standing support of the RCSB PDB, which has served as a model for managing "big data" and making it accessible to a broad and diverse community of users. The Committee was thus particularly gratified by comments from the NSF representative regarding their increased recognition of the value of long-term support for databases like the RCSB PDB.

Responses to 2014 RCSB PDB AC Recommendations

- PDBAC: Pursue funding to develop approaches for supporting data from integrative/hybrid methods
  Response: Proposals submitted.

- PDBAC: Terminate the legacy deposition system (ADIT)
  Response: ADIT retired July 2015 for x-ray crystal structures

- PDBAC: Continue to provide mobile-friendly services
  Response: Redesign of Structure Summary and PDB-101 pages to respond to display type.

- PDBAC: Develop a focused Education Plan
  Response: Comprehensive redesign; described below.

- PDBAC: Make more information available on unpublished structures
  Response: Requires further discussion with wwPDB and community stakeholders.

.

PDB Metrics

In aggregate, 10364 depositions were processed between January 1st and December 31st 2014 with a two-week average turnaround, a decrease from the 10566 entries deposited in 2013. Based upon the number of entries deposited in 2015 to date, it is estimated that 11000 entries will be deposited in 2015.

Breakdown of depositions by discipline in calendar 2014 was as follows:

X-ray:	9586 (93% of entries deposited, down from 9697 in 2013)
NMR:	515 (5%, down from 590 in 2013)
EM:	240 (2%, up from 234 in 2013)
Other:	23 (0.3%, down from 45 in 2013)

Breakdown of depositions by wwPDB processing site in calendar 2014 was as follows:

RCSB PDB:	6040 (58%)
PDBj:	1779 (17%)
PDBe-EBI:	2545 (25%)

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

North America	37%
Europe	33%
Asia	19%
Industry	7%
South America	<1%
Australasia	4%
Africa	<1%

During 2014, RCSB PDB's website at http://rcsb.org was visited each month by an average of 283,358 unique visitors and 668,348 unique visits. A total of 25.033 GB of data were accessed.

During the same time period, more than 505 million data files were downloaded from the PDB archive *via* the wwPDB member FTP and websites (RCSB PDB: 347,283,931; PDBe: 100,393,784; PDBj: 57,683,377).

2015 RCSB PDB AC Discussion

Integrative/Hybrid Methods

Dr. Helen Berman presented an overview of how the RCSB PDB is meeting the new challenges presented by deposition of structures determined by multiple experimental methods. Berman summarized the discussions held at the Hybrid Methods Task Force meeting at EMBL-EBI in Hinxton, UK in October 2014. The resulting set of recommendations, which were published in *Structure* in July 2015, identified issues regarding model and data archiving, structure representation, validation, and publication standards to be dealt with by all the wwPDB partners. In addition, a federation of model and data archives, including the newly-formed Small-Angle Scattering Biological Data Bank (SASBDB), will be established to handle depositions and create a single hybrid model repository. A Working Group led by Berman and Advisory Committee members Trewhella, Sali and Schwede are leading a Task Force and subgroups that are grappling with these issues and confer monthly to discuss progress and coordinate efforts. The Committee was gratified to hear that an NSF EAGER grant has been obtained to support some of these new efforts and that a new proposal on hybrid model validation has been submitted to the NSF. The Committee fully supports the RCSB PDB efforts in this critically important new area in structural biology and hopes that the necessary additional funding will be forthcoming.

Data In: Deposition, Annotation, and Quality Assessment

Dr. John Westbrook gave an overview of the activities of the curators and developers who manage data deposition and annotation. The team does an impressive job of curating data and developing tools for submission and curation, thanks to their breadth of expertise in x-ray crystallography, NMR, EM, small molecules, software and statistics.

Dr. Jasmine Young provided an update on depositions, which during 2015 transitioned to exclusive use of the Common Deposition & Annotation System (D&A), with phase-out of the older ADIT system over the period January – June 2015. The new D&A system has made possible an impressive increase in throughput, with approximately 50 entries per month processed by each full-time employee (FTE). This enabled the RCSB PDB to handle over 6,000 entries over the past year. These entries are of increasing complexity and size, which can now be handled efficiently, thanks to the adoption of the PDBx format. Young also updated the Committee on numerous improvements to biocuration, including annotation of chimeric protein sequences, improved ligand annotation and better workflow management, which is improving both the user experience and increasing curation efficiency. The Committee views both of these as mission-critical to the long-term ability of the PDB to serve both depositors and users, the latter of which are increasingly non-experts. The remarkable decrease in processing time, which has decreased from a median of 16.5 days with ADIT to 1.6 days with the new D&A tool, has, however, had unintended consequences, namely a large increase (~150%) in the rate at which some users replace coordinates each month, presumably in response to the results of validation reports. The Committee felt that, while improvements to structures are to be welcomed by the community, it will be important to reduce the replacement rate to ensure long-term productivity

and throughput. The Committee recommends that the RCSB PDB investigate the reasons for coordinate replacements, and experiment with incentivizing researchers to validate their data prior to depositing coordinates and to correct coordinate errors prior to deposition. As part of this effort, the Committee recommends that the RCSB make the PDB validation software available to users and developers via a web-accessible interface as well as for download, for those who wish to install a local version.

Dr. John Westbrook informed the Committee on the deployment of the D&A tool, done in partnership with the wwPDB and implemented over the period January 2014 to September 2015. The Committee was gratified to hear that a secondary site for disaster recovery was set up in April 2015 in partnership with the wwPDB and feels this was a critically important measure. Dr. Westbrook also updated the committee on further developments in data standards, guided by recommendations of the PDBx/mmCIF Working Group chaired by PDB AC member, Dr. Paul Adams. Changes to be implemented include the NMR Exchange Format (NEF) for restraint data and external references files (ERFs) such as links to the Cambridge Structural Database. Looking ahead to 2016, the Committee endorses the plan to remediate carbohydrates, residual B factors and space group settings, as previously discussed. The Committee also looks forward to completion of version 2.0 of the D&A tool, which is being developed in partnership with PDBe. The Committee is deeply concerned that more than one year has passed since the original completion deadline, because the outstanding NMR and 3DEM deposition functionalities have not been completed. The timely completion and implementation of D&A 2.0 is of paramount importance to the long-term ability of the wwPDB to meet its obligations to the larger community. The Committee thanks Dr. Jasmine Young for her willingness to help manage the project and looks forward to a rollout in early 2016.

Data Out: Data Access and Exploration

Dr. Peter Rose introduced the newly designed Structure Summary page, which had become cluttered and difficult to navigate as features were incrementally added. The Committee found the new design, whose look and feel was based on last year's redesign of the main PDB page, to be clean and user-friendly, making recently added features such the Protein Feature View and structure visualization easier to access and use. The Committee was pleased to hear that this feature has also been made accessible on mobile devices, which last year constituted 10% of web traffic and are increasingly being used to access the PDB. Dr. Andreas Prlić showed the Committee how to access mutation information in Protein Feature View as well as graphical summaries of structure validation and the web-based 3D structure viewer. These features are thoughtfully designed and easy to use. The importance of providing multiple tools accessible to naïve users was driven home by the fact 75% of RCSB PDB users are non-specialists. The Committee commends the RCSB PDB for its ongoing comprehensive use of web analytics to measure usage and analyze user demographics. These data are important for the ongoing ability of the PDB to meet the needs of its user community and will be critical for making their case to funding agencies. At the same time, the RCSB has also made wise use of user communications with the Help Desk to obtain feedback and identify areas for improvement. Looking ahead, the Committee endorses the RCSB PDB plans for further improvements, including visualization of ligand electron density. The Committee also supports the proposal to map structures to protein families and to biological pathways, which will be highly useful to the general user community.

Education and Outreach

The RCSB PDB has long carried out an impressive array of outreach and education activities. Last year, the Committee expressed the concern that these efforts needed to be more focused in order to stay within the current budget constraints while maximizing impact. The Committee was highly impressed by the outstanding new education and outreach plan presented by Dr. Shuchismita Dutta. The new plan, a comprehensive and thoughtful restructuring of education efforts, builds upon successful elements of previous efforts. The partnerships with educators to develop teaching materials and use of HIV/AIDS and diabetes, as frameworks to educate students at various levels about biomolecular structure, are all highly attractive elements. The overall plan for developing curriculum modules and then field-testing and assessing their impact is well thought-out and focused. While currently aimed at high school and college students, the plan to extend the reach to healthcare professionals and continuing medical education could further broaden the impact. As additional materials are developed, the Committee expects that the RCSB PDB will determine how these can best be publicized and made readily visible on the web site.

One of the most popular education and outreach resources on the RCSB PDB web site is PDB-101, which provides a variety of educational materials that are utilized by students and faculty alike. Dr. David Goodsell provided the Committee with a preview of the redesigned PDB-101 web interface, which addresses some shortcomings of the current version while making the site more user-friendly and easier to update. Improvements include the ability to search the popular Molecule of the Month pages, which were previously accessible through a menu only, as well as clearer and more intuitive menus and organization. Given the popularity of PDB-101, which accounted for an impressive 12% of RCSB PDB web traffic in the past year, the Committee expects that these changes will maximize the utility of these features and looks forward to the planned rollout at the end of the year.

Management

Dr. Stephen Burley provided a summary of the RCSB PDB organization, current funding and responses to NSF requirement for the current funding period. The Committee once again commends Drs. Burley and Berman for ensuring a seamless leadership transition last year and for continuing to work as an effective team with the help of Deputy Director Christine Zardecki. Burley is currently also directing the UCSD site but hopes to recruit a replacement for Dr. Phil Bourne, who left for the NIH last year. The RCSB PDB has had impressive success in obtaining grants for specific outreach and technology development projects, in addition to its core support from the NSF/NIH/DOE. As stipulated in the NSF requirements for the current 2014-2018 funding period, the RCSB PDB has developed a business model, diversity plan, and assessment plan and has revised the guidelines for membership of the advisory committee. The plan to appoint members to 3-year renewable terms will ensure turnover and strengthen the ability of the RCSB PDB to appoint members in new areas, as witnessed by the addition this year of members with expertise in hybrid methods, cryoEM and visualization, as well as representatives from industry. The Committee Chair has agreed to stay on through the next major grant renewal and will be replaced in 2019. To ensure sustainability in future years, the RCSB has been able to dramatically increase efficiency and rebalance ‘Data In’ tasks, thanks to the automated D&A tool. Plans to extend the wwPDB franchise to China and India will enable the PDB to meet expected increased demands from investigators throughout Asia.

Plans for financial support

The RCSB is currently on solid financial footing, thanks to success in obtaining grants for targeted projects. Continued success in this arena, together with plans to seek private and corporate funding, will be important for implementing plans for additional activities.

There was a discussion with Dr. Peter McCartney of the NSF, who participated via telephone. Dr. McCartney told the committee that the NSF views the RCSB as a major resource that has been able to maintain support and a positive profile at the NSF because of its well-defined scope. Dr. McCartney praised the RCSB PDB for maintaining this focus and recognizing "what it is and what it isn't." The Committee was very gratified to hear from Dr. McCartney that the NSF recognizes the value of providing long-term financial support to databases like the PDB. This is a shift from the previous funding philosophy, which considered NSF support to be seed funding, and is enthusiastically welcomed by the Committee. In the discussion following the telephone conversation, the Committee and the RCSB leadership agreed that it would be beneficial to include representatives from the NIH and DOE in next year's conversation with the NSF, with an eye towards planning for what is likely to be a competing renewal in 2018.

Matters arising

The Committee was asked to provide input on a number of matters confronting the RCSB PDB. The PDB leadership solicited advice on a proposal to enable three-dimensional visualization of the structural impact of coding SNPs (single nucleotide polymorphisms) and other genetic variations. While the Committee agreed that this could, in principal, be of great utility to the broader biological community, particularly those lacking expertise in structural biology, the issue generated a wide ranging discussion of how such a plan would be implemented, what the focus would be, who would carry out the project and how it could be ensured that meaningful models would be generated. The Committee recommends exploring the issue with a pilot project, perhaps focusing on disease-causing mutations and to measure interest in the research and educational communities before broadening the project's scope. The Committee endorsed the RCSB PDB's proposal to develop resources for exploring protein families and pathways and recommends also making available precomputed structure superpositions for families of related proteins.

The Committee discussed at length the question of what the RCSB PDB, or the wwPDB, should do about errors that are detected by validation software, biocurators or users, but are not corrected by the author. The Committee endorses the proposal to put a comment section on the RCSB structure summary page where comments from annotators could be posted and responses from depositors could be solicited. The Committee also supports the proposal to identify whether there are journals whose reported structures have particularly elevated rates of problematic structures and consider engaging editors in the effort to increase author responsiveness to queries from curators.