RCSB Protein Data Bank Advisory Committee
Report of May 8, 2019 Annual Meeting
In person and teleconference

Chair: Paul Adams

Membership:

Present: Paul Adams, Judith Blake, R. Andrew Byrd, Wah Chiu, Kirk Clark, Roland L. Dunbrack, Jr., Thomas E. Ferrin, Sue Rhee, Jill Trewhella

Virtual: Paul Craig, Catherine E. Peishoff

Absent: Peter Andolfatto, Bridget Carragher, Robert B. Darnell, Paul Falkowski, Mandë Hollford, Torsten Schwede

Funding representatives: R. John Knowlton (NCI), Ward Smith (NIGMS), Ramana Madupu (DOE), Peter McCartney (NSF), Amy Swain (DOE)

RCSB PDB AC E-mail Addresses:

PDAdams@LBL.gov, pa2543@columbia.edu, Judith.Blake@JAX.org, bcarr@nysbc.org, byrda@mail.nih.gov, wahc@stanford.edu, kirk.clark@novartis.com, paul.craig@rit.edu, Robert.Darnell@Rockefeller.edu, roland.dunbrack@fccc.edu, falko@marine.rutgers.edu, tef@cgl.ucsf.edu, mholford@hunter.cuny.edu, peishoffc@gmail.com, srhee@carnegiescience.edu, torsten.schwede@unibas.ch, j.trewhella@mmb.usyd.edu.au

RCSB PDB Leadership: Stephen Burley (Director), Helen Berman (Director Emerita), Andrej Sali (UCSF Site Head)

RCSB PDB Leadership E-mail Addresses:

sburley@proteomics.rutgers.edu, berman@rcsb.rutgers.edu, sali@salilab.org

Executive Summary

The Advisory Committee (AC) to the Research Collaboratory for Structural Bioinformatics (RCSB) held a meeting on 8 May 2019 to consider management and enhancement of the Protein Data Bank (PDB).

Agenda items included

- Welcome/Introductions/State of the RCSB PDB
- Deposition/Biocuration (Service 1)
- Funding: Current Status; Funder Discussion
- Management
- Mol* 3D Visualization Demonstration
- I/HM Progress/Preview on Validation
- Building Next Generation Archive Management/Access (Service 2) and Data Exploration (Service 3)
- Outreach/Education (Service 4)

- Discussion/Feedback on Growing the RCSB PDB Data Consumer Community by Delivering New Tools for RCSB.org and Integrating with Additional Data Resources

The meeting was opened by Dr. Stephen Burley. Other RCSB PDB participants were Helen M. Berman, Robert Lowe, John Westbrook, Jasmine Young, Christine Zardecki (Rutgers); Andrej Sali (UCSF); and Jose Duarte, Alex Rose, and Joan Seguro (UCSD). Appendix 1 provides a summary of the RCSB responses to the 2018 Advisory Panel meeting recommendations. Appendix 2 provides a summary of global PDB deposition and data access statistics in 2018.

Overall Comments from the Advisory Panel

The team is congratulated on an excellent set of presentations, and their collective leadership of the RCSB and the individual projects over the last year. The committee recognizes the efforts of the RCSB Director, Stephen Burley, whose leadership has had, and will continue to have, a very positive impact on the team, with staff taking on more responsibility and developing their project management and presentation skills. The committee was extremely happy to receive the positive news about NSF/NIH/DOE funding, while recognizing that the lack of full funding presents problems in executing on plans that are of great significance to the scientific community. The team, and Stephen in particular, are congratulated on their efforts, which have led to the first increase in funding in a decade.

It is recognized that planning for further funding to close the gap between funds requested and awarded is a major effort, which the RCSB team is working hard to achieve. The committee were concerned that this is placing an extreme burden on the RCSB team, and we encourage the Director to seek ways to help reduce this burden. These may include staff development to enable independent submission of proposals, and engagement with the Rutgers Development Office, and their Public Relations Office. Alternatively, maybe there are opportunities to recruit existing tenure track faculty to the program more formally, to assist in fundraising. The inclusion of UCSF, under the leadership of Andrej Sail, as a formal RCSB partner is a very positive step, in particular for the development of hybrid methods.

Finally, although plans weren't presented in detail, the committee was very heartened to hear that proposals are being submitted that embrace new research directions such as machine learning. The RCSB is well positioned to have an impact on structural biology by exploiting these new approaches.

Recommendations for future meetings

- Limit presentations to 15-20 minutes each, leaving adequate time for discussion. The content of the presentations was very thorough and helpful, but presenters need to be succinct in speaking to the content.

- It would be helpful for the advisory committee if the speakers could each provide a concise set of questions or topics for discussion at the end of their presentation. This would help move the discussions in directions most beneficial to the RCSB.

Detailed Advisory Panel Comments and Recommendations

Deposition/Biocuration

Dr. Jasmine Young is congratulated on her efforts leading this team. She has had a transformative impact on the OneDep project, demonstrating excellent leadership skills. The efforts Jasmine is putting into project management and resource planning are clearly paying off, and are strongly encouraged as a

general practice across the RCSB. The committee was impressed with the turn around of the ligand validation project. Jasmine has been very effective in transferring the project from PDBe, where staff changes were hindering progress. The donation of code from Global Phasing for ligand validation is also a very positive step forward.

The committee was pleased to learn that there are wwPDB-wide annotation protocols, which are publicly available. However, it is clear that not every situation is currently addressed, with a good example being antibodies. The problems with metadata are significant enough that a remediation will be required. The committee fully endorses the RCSB taking a leadership role in those future efforts. Carbohydrates have been a perennial problem in the PDB, so the efforts towards remediating carbohydrate entries are positive and long overdue. The deposition of cryo-EM structures, in particular different conformational states will be an increasing challenge, which the RCSB should engage the broader community to address.

Recommendations

- The team is encouraged to factor potential growth in experimental techniques, particularly cryo-EM into their deposition projections

- The transition to mmCIF for crystallographic structure deposition in July 2019 is an excellent opportunity for the wwPDB as a whole. If successfully managed it can be expected to lead to a further increase in annotator efficiency. The team is encouraged to closely monitor the transition and work with community software developers to ensure that users have the tools for effective mmCIF deposition, and that the RCSB staff are prepared for first wave in July.

- The RCSB should consider providing visual views of the ligand validation information analogous to the wwPDB validation summary graphic for macromolecular structures on the structure summary pages - so that less knowledgeable users can easily see if there are issues with ligands.

- The RCSB is encouraged to work with the carbohydrate community to provide open source software tools to handle the new carbohydrate description mechanisms - this will greatly increase the likelihood of other community software packages adopting these mechanisms.

Archive Management/Access and Data Exploration

The committee heard about the ongoing efforts to replace the underlying “data out” computing infrastructure at the RCSB from Dr. John Westbrook. This is clearly an essential activity, as the current technology is now somewhat out-of-date. Committee members were impressed with the state of the new implementation (at search.rcsb.org), with huge improvements in performance being obvious. There was some discussion about whether the computing infrastructure and associated data storage is an activity that needs to be undertaken by the RCSB, given that there are many enterprise systems worldwide that are built on similar technologies. The RCSB should consider investigating whether a transition to third party hosting in the future is practical and/or cost-effective in today’s IT landscape. Maybe there are also possibilities for deeper engagement with data science programs at the RCSB institutions.

A substantial part of the presentation and subsequent discussion focused on the search features in the new system. There are clearly some very positive developments, which are enabled by the new underlying technology, including better browsing and text searching. However, there was an impression that the activities are mainly focused on replicating the current search functionalities. While this has value, the committee felt that the RCSB might be missing a big opportunity to rethink the whole search concept.

It was noted that many of the search functions primarily help structural biologists retrieve atomic models. Given the large number of non-structural biology users of the site, some thought should go into devising new ways to help them discover structures and present information in more meaningful ways. Some committee members were quite excited by the possibilities of the new 3D search feature, whereas others did not see that a clear use case had been developed with RCSB users in mind.

Dr. Westbrook provided information about data security at the RCSB in response to a committee question. The efforts were described as part of a huge and very impressive effort to obtain certification for the RCSB. The team is applauded for the work that went into this certification, and for being leaders in this area for academic enterprises.

Recommendations

- The RCSB is encouraged to think about how new search features, and/or modification of existing approaches could help serve the broader community of visitors to the site. In particular ways to provide data out that are more meaningful to those not familiar with 3D structures.

- The team should develop a firm timeline for delivery of the new data out technology to the user community.

Outreach/Education

The committee heard about the outreach and education efforts over the last year, with some highlights being the focus on antimicrobial resistance, and a number of in person outreach activities. These outreach and education team is clearly working very well under Dr. Christine Zardecki's able leadership. The proposed focus on Drugs and the Brain in the coming year is strongly supported by the committee, as this is likely to have broad appeal to the public. A number of mechanisms to increase outreach and community engagement were discussed, including enabling the production of promotional items, further involvement in the development of curricula, and the creation of games that would expose younger people to structural biology.

Recommendations

- The RCSB should consider making links to 3D printer files, along with instructions for printing more readily available from the RCSB website, as this might be of increasing interest to the community.

Hybrid Methods/Visualization

The plans for developing a hybrid methods validation pipeline were presented by the UCSF lead, Andrej Sali. It was encouraging to see that there is now NSF support for some of these activities (development of the pipeline), with core RCSB support covering the remainder (implementation in OneDep). The proposed approach for hybrid methods validation stems from a recent workshop held in Baltimore. The team is applauded for seeking community input to help guide these efforts. Helen Berman is also recognized for her fund raising efforts to support the workshop, and her continued engagement in hybrid methods development through the NSF funding. Her continued participation in this project is viewed as key to its success. The committee expressed some concern about how some of the hybrid methods data would be validated, with cryo-electron tomography reconstructions being a particular issue. The RCSB approach of developing a pipeline that can run a broad range of hybrid method models/data, albeit imperfectly at the beginning, is appropriate - recognizing that "perfect is the enemy of good". The

proposed use of Bayesian approaches for validation in the future is very exciting, and has a good statistical basis. The RCSB is encouraged to consider this a topic for future fund raising.

The committee heard from Axel Rose about the development of a new web-based molecular visualization system, called Mol*. This is a collaborative effort with PDBe, which will ultimately replace the existing NGL and LiteMol software (from RSCB and PDBe respectively). The need for a new system wasn't well articulated, but the demonstration showed that it is capable of complex molecular representations. Some committee members expressed the concern that the current Mol* software is very hard to use, which was acknowledged by the RCSB team. Efforts are underway to create a user-friendly GUI. Developing use case examples for the new viewer would provide a means to better learn how to best use it. Maybe this would be a good project for high school students?

Recommendations

- The NGL library has been used by a few groups to develop new tools. The RCSB are encouraged to develop guidelines to help those groups transition to the Mol* system.

Management

The committee applauds the RCSB's succession planning efforts. Only too often these are overlooked in an academic setting, leading to challenges during transitions. There were some concerns that the succession planning for the Data & Software Architect Lead appeared to exclude consideration of the potential for internal candidates. The RCSB should recognize that this is potentially an opportunity to develop internal staff for the role. The potential for advancement to such a role is motivational, and could play a role in attracting high potential recruits. In all cases it is assumed that an open search will be conducted.

The establishment of a PDBc (in China) is very encouraging. This group would be an associate to begin with, then move into Core if qualified. The committee is very supportive of this activity and the approach. There were some questions about barriers to interactions with foreign entities in the current climate, that RCSB leadership should discuss with federal funders. The advisory committee did recognize that the RCSB's aspirations for funding growth, and the inclusion of other centers such as PDBc, will significantly increase the management workload for the Director and his team. This increase needs to be planned for. The committee was impressed by the level of project planning within RCSB projects, and acknowledges that for some cases, this spans multiple institutions internationally. There were some concerns about how projects were prioritized.

Recommendations

- The RCSB leadership should develop a management plan to deal with increased activity that will require funding and additional wwPDB centers.

- The RCSB should provide information to the committee at the next advisory meeting about how projects are prioritized across the center.

Funding

The committee heard about the need to raise significant additional funding, and it may be that such funding could have the potential to offset the funding allocated from federal agencies (NSF/NIH/DOE). Ensuring any such offset is managed so that funds freed up remain available to RSCB for currently

unfunded activities will require careful communication and management with the agencies. While the current federal support is significant and welcomed most enthusiastically, the shortfall is clearly recognized. It should also be recognized that federal budgets for supporting the RCSB activities are constrained and creative mechanisms for new funding will be required. The committee were very pleased that representatives of NSF, NIH and DOE were able to attend the advisory meeting remotely, and greatly appreciated the discussion session with Peter McCartney.

There were many options for fundraising that were discussed during the meeting:

- Although HHMI hasn't been successful, some committee members wondered if Janelia Farm might provide some opportunities given their recent call for ideas about their future directions.
- Philanthropic support is very challenging to obtain, but maybe it would be good to target prior supporters of science like Jim Simmons.
- The recent NCI ITCR program might provide some opportunities.

Recommendations

- A direct discussion with Susan Gregurick at NIH might help understand the landscape for upcoming data-related opportunities.
- RCSB should develop, immediately, 1-2 pages documents describing potential projects (outside their core mission) that will enhance the RCSB activities. These will be invaluable for seeking funding from industrial sources.

Expanding the User Base/Measuring Impact

There was substantial discussion about the desire to increase the RCSB user base, as an indicator of the need for continued and increased funding. This is a reasonable strategy; however, the committee suggests that the RCSB develop a more considered justification for this activity.

The committee suggested multiple disciplines where growth in the use of structural data would likely be possible and scientifically advantageous: Neurobiology, Bioengineering, Synthetic biology, Evolutionary biology, Environmental biology, Pharmacology and Artificial Intelligence. There were also some specific projects that might be targets for engagement including the Human Cell Atlas and the Plant Cell Atlas. The committee also felt that engagement with other larger biological data resources would provide routes to new users. In particular the Alliance of Genome Resources might be a natural partner, with the potential for joint proposals for funding. It was also felt that the time was right for a renewed interaction with NCBI and the NLM, with the goal of promoting structural biology in NCBI resources.

Overall the committee thought that multiple strategies will need to be followed to increase the user base, and that those strategies will need to be well coordinated. Fundamentally, it is unlikely that researchers not already acquainted with structural biology will become new visitors to the RCSB web site. Therefore, the RCSB will need to seek ways to partner with other groups to link from their web sources to RCSB data. Low hanging fruit might be providing easy mechanisms for online publications to link to structural data (PubMed already does this, with links to NCBI resources). It was noted by the committee that undergraduate educators view themselves strictly as consumers, rather than contributors, and that there is an opportunity to change that. Maybe there are mechanisms that can be offered to encourage this group to contribute, or even become advocates for the RCSB and structural biology in general. Expanding

the user base into non-structural domains will require careful thought about how to serve information in ways that are of interest in those domains.

Recommendations

- The RCSB should develop a concise justification for the need for an expanded user base, and the general topic of demonstrating impact, with clear goals.
- The 50th anniversary of the PDB in 2022 provides a great opportunity for community outreach. The RCSB is encouraged to plan for high profile coverage of the event, including the national press. Examples discussed included special issues of journals, highlight articles in Science or Nature, CBS Sunday Morning, family days at NIH, a BBC special (online or broadcast), and engagement with educational networks, e.g. a Nova special.
- The RCSB should investigate opportunities for partnership with social sciences academic groups with expertise in researching user engagement and outreach. There may also be opportunities for funding projects to help in this area.
- The RCSB has an opportunity to lead in the promotion of structural biology in science at a national level. This is a theme that could gain momentum if pursued in collaboration with other interested groups but will likely require engagement with federal officials and congress.

Appendix 1: Responses to 2018 RCSB PDB AC Recommendations

The AC was provided with the following responses:

PDBAC:

on Funding: The AC therefore supports the RCSB PDB leadership in its efforts to raise additional funds from pharmaceutical and biotech companies, as well as the Howard Hughes Medical Institute, to support critical projects that are not funded by the current grant. The AC recommends that the RCSB PDB develop specific project proposals that could be supported by one or more of these entities.

Response:

To be discussed in Funding; we are actively trying to fundraise and develop proposals to support 3DEM, XFEL, and I/HM research.

Evolving Experimental Methods: 3DEM
Project proposed to HHMI–Rejected
Project proposed to ThermoFisher–under review
Evolving Experimental Methods: XFEL
Project proposed to NSF–proposal in preparation
Emerging Integrative/Hybrid Methods
Project proposed to ThermoFisher–under review

PDBAC:

on Federation of Data Resources: The new organizational plan of the wwPDB … as federated resources, presents new opportunities and challenges that will require new technical and financial support. It will therefore be essential for the RCSB to obtain additional grant funding, as well as support from industry and non-profits…

Response: Action deferred until EMDB joins wwPDB

PDBAC:

on NIH Funding: Given the impact of the RCSB PDB on a broad range of NIH institutes, the AC expressed the hope that future funding could come from the Office of the Director rather than just from the budgets of NIGMS and NCI.

Response:

To be discussed in Funding; we will continue to communicate and demonstrate the impact of the PDB across the NIH.

PDBAC: on Support for growth in cryo EM, integrative structural biology, and SFX/XFEL:

The AC recommends that the RCSB PDB leadership develop specific project proposals that are more detailed than the three currently outlined core challenges, and to produce materials that articulate these projects. In addition to assisting the RCSB leadership in raising funds, these proposals could help AC members seize opportunities for fundraising.

Response: To be discussed;

Would one page summaries be beneficial?

- Problem statement
- Key deliverables
- Required FTEs
- Timeline
- Cost

Appendix 2: PDB 2018 Metrics

In aggregate, 12,179 depositions were received and processed between January 1st and December 31st, 2018, with an average turnaround of two weeks. This represents a decrease from the 13,049 entries deposited in 2017. Based upon the number of entries deposited in 2019 to date, it is estimated that 13,603 entries will be deposited this year.

Breakdown of depositions by discipline in calendar 2018 was as follows:

- X-ray: 10,594 (87% of entries deposited, down from 11,889 in 2017)
- NMR: 418 (3.4%, down from 460 in 2017)
- EM: 1,140 (9.4%, up from 658 in 2017)
- Other: 27 (.2%, down from 42 in 2017)

Breakdown of depositions by wwPDB processing site in calendar 2018 was as follows:

- RCSB PDB: 5116 (42%)
- PDBj: 2897 (24%)
- PDBe-EBI: 4166 (34%)

Breakdown of depositors by location in calendar 2018 was as follows:

- North America 34%
- Europe 33%
- Asia 23%
- Commercial 6%
- South America 1%
- Oceania 3%
- Africa <1%

During 2018, RCSB PDB's website at http://rcsb.org was visited by millions of unique visitors.

During the same period, an estimated 749,356,769 679 million data files were downloaded from the PDB archive via the wwPDB member FTP and websites (RCSB PDB: 500,921,718; PDBj: 66,244,841; PDBe: data for 2018 unavailable due to GDPR (actual data provided for other years).