RCSB Protein Data Bank Advisory Committee

Report of March 15th, 2022 Annual Meeting

Teleconference

Chair: Paul Adams

Membership:

Present: Paul Adams, Peter Andolfatto, Bridget Carragher, Wah Chiu (after 1pm), Kirk Clark, Roland Dunbrack, Paul Falkowski, Thomas Ferrin, Mandë Holford, Cathy Peishoff, Sue Rhee (joined late), Torsten Schwede, Lance Stewart, Kevin H. Gardner, Takita F. Sumter, Jill Trewhella

Absent: Robert B. Darnell

RCSB PDB AC E-mail Addresses:

Bridget Carragher <bcarr@nysbc.org>, darnelr@rockefeller.edu, falko@marine.rutgers.edu, Jill Trewhella <jill.trewhella@sydney.edu.au>, kgardner@gc.cuny.edu, kirk.clark@novartis.com, Lance Stewart <ljs5@uw.edu>, mholford@hunter.cuny.edu, pa2543@columbia.edu, Paul Adams <pdadams@lbl.gov>, peishoffc@gmail.com, roland.dunbrack@gmail.com, Sue Rhee <srhee@carnegiescience.edu>, sumtert@winthrop.edu, Tom Ferrin <tef@cgl.ucsf.edu>, Torsten Schwede <torsten.schwede@unibas.ch>, Wah Chiu <wahc@stanford.edu>

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 virtual meeting on March 15th, 2022 to review recent progress and provide feedback on specific questions.

Agenda items included

- Welcome and Introductions
- State of the RCSB PDB
- Deposition/Validation/Biocuration, Remediation, and Archive Management
- Data Exploration (RCSB.org)
- Outreach/Education (PDB-101); PDB50
- Operations and Funding
- Strategic Initiatives
- Discussion

The meeting was opened by Dr. Stephen Burley. Other RCSB PDB participants were Helen M. Berman, Jasmine Young, Yana Rose (UCSD), Christine Zardecki (Rutgers), and Andrej Sali (UCSF). Also, in attendance for parts of the meeting were representatives of funding agencies: Steven Ellis (NSF), Paula Flicker (NIH-NIGMS), Jerry Li (NIH-NCI), Ramana Madupu (DOE), and Amy Swain (DOE).

Appendix 1 provides a summary of the RCSB responses to the 2021 Advisory Panel meeting recommendations. Appendix 2 provides a summary of global PDB deposition and data access statistics in 2021.

Overall Comments from the Advisory Panel

The team is again congratulated on their outstanding work over the last 12 months, as the field of structural biology undergoes significant changes with the wider availability of highly accurate structure prediction methods. The team gave an excellent set of focused presentations, and their collective leadership of the RCSB and their individual projects is applauded. It was good to see that new people have joined the team despite the pandemic, and that interviews are underway for additional hires. The committee recognizes that the major activity for the team in the coming months is preparation for the renewal of the RCSB funding.

Recommendations

- The committee encourages the RCSB team to make use of the committee's expertise to assist in preparing for the renewal.

Detailed Advisory Panel Comments and Feedback

The committee welcomed new members, Kevin Gardner and Takita F. Sumter, and look forward to working with them in future meetings. The addition of Kevin Gardner is important as he replaces long serving member Andy Byrd's expertise in NMR after his retirement from the committee last year. The committee also thanked recently retired members Judy Blake and Paul Craig for their service. Paul's current sabbatical with RCSB PDB was noted as a very positive development. The committee also thanked current member Jill Trewhella who will be retiring from the committee this year.

The committee is greatly saddened by the unexpected passing of John Westbrook last year. He was a valued and close colleague for many of us, for many years. It is clear that this is a huge loss for the RCSB and the research community.

Deposition/Validation/Biocuration, Remediation, and Archive Management

The committee heard from Jasmine Young about deposition and biocuration activities. We continue to be very impressed with Jasmine's leadership in this area, and the report on last year's activities was very positive. There have been numerous improvements to the OneDep system, and the level of automation continues to increase, with the result that 76% of new entries now pass automatically through significant parts of the process. The committee wondered what it would take to further increase the level of automation. We heard that the RCSB is focusing on gains for the more routine, simpler structures. Review of the average deposition times over the last few years suggests that maybe the limit has been reached for the routine structures, and more emphasis will need to be placed on the complex structures and what gains in efficiency can be achieved there. This will become a larger fraction of depositions as the number of cryo-EM depositions increases. The improvements in cryo-EM validation are therefore very welcome, and clearly an area for future developments. The committee was pleased to hear that the assemblies will be available soon - within 2 months, and that SIFTS will be included in the next version of the archive as part of a collaborative grant with the PDBe team. We were asked if we had any concerns about the Deposition/Biocuration work underway with the wwPDB partners. We don't and again congratulate Jasmine for her leadership of OneDep and creating a strong collaboration across the wwPDB.

Recommendations

- At a future committee meeting the team should present their strategy for increasing automation of deposition of more complex structures, including any implications for developer resources needed. This might be timely given the renewal of funding activities in the coming months.

Data Exploration (RCSB.org)

The committee heard from Yana Rose on the improvements in data delivery through the RCSB.org web site. We recognize that she has stepped in to fill some of the void left by John's passing, and the committee congratulates her on her achievements already. We were collectively very pleased to see significant improvements in search functionality, integration with other data resources, and new data visualization tools. The ligand validation information is excellent and of great importance to the user community. There was some discussion about increases in the number of ligand-containing depositions, with the large number of SARS-CoV-2 fragment campaigns providing jumps in both 2020 and 2021. This was followed by a discussion about the number of ligand structures that might be needed to solve the protein/ligand prediction problem with machine learning. A figure of 50,000 ligands was postulated, which will eventually be in reach. It was suggested that encouraging Pharma to deposit all their older structures might help, although it is appreciated that a significant amount of work would be required to complete these structures to deposition quality. An alternative would be a concerted academic effort to systematically solve protein/ligand complexes, analogous to the PSI effort that provided many of the structures that helped solve the protein structure prediction problem. In addition, it may be possible to conduct machine learning with protein/ligand complexes made available by Pharma in a federated learning approach resembling a pre-competitive consortium model.

Looking to the future of the RCSB.org web site, the committee is very enthusiastic about the recently initiated Rutgers user experience design review and the involvement of Paul Craig and the BioMolViz community. We expect that this will further enhance the usability of the site in the future. We also heard some of the anticipated plans for the new PDBc site. It is likely that they will serve a Mandarin language data out site, principally for users in China. It is not clear if this will have any significant impact on the number of users accessing the RCSB.org site.

Recommendations

- Report out to the committee on the results of the UXD review at the next meeting, and seek our input prior to implementing the proposed changes.

Outreach/Education (PDB-101); PDB50

Christine Zardecki presented many of the great efforts in outreach and education over the last year. A highlight was the activities celebrating 50 years of the PDB archive, which included 8 virtual symposia and webinars, special journal collections and publications, and a new PDB50 section at the RCSB.org website. These were accompanied by several other special projects, including mention in the Congressional Record. The committee were pleased to see that the COVID-19 outreach efforts have continued, including education activities through Rutgers. The efforts to promote student engagement in science communication and their collaboration on the Molecule of the Month are recognized and strongly encouraged. The committee wondered what else can be done to extend the reach to underserved communities, appreciating that some of those activities are better in person than virtual.

Recommendations

- At the next committee meeting present progress and future plans for outreach to underrepresented minorities and underserved communities.

Operations, Funding and Strategic Initiatives

The committee heard about recent operational improvements, recruiting activities and diversity, equity, and inclusion. The committee was very pleased to hear about the NSF-funded upgrade to the computing infrastructure, which should accommodate anticipated growth through 2025. We were also pleased to see that a number of positions are being actively recruited and that across the RCSB sites there is active engagement in DEI principles and leveraging of programs at the participating institutions.

Two strategic initiatives were highlighted, both very important. The migration of backend RCSB.org services to cloud computing is of great significance, and essential for the feasibility of RCSB operations in the future. The committee thanks NIGMS for their containerization supplement and applauds Stephen Burley for working with Amazon to obtain significant storage capacity as part of the AWS Open Data Sponsorship Program. The committee wholeheartedly supports these data and compute migration activities, and believe they can provide a model for other large data projects.

The second initiative was the PDB-Dev archive and improved validation information for integrative/hybrid methods structures. This is an important growth area, which will likely be accelerated by the availability of the computed structure models. There was some discussion of how they might be used by the PDB-Dev archive, and the committee is pleased to hear that this archive will merge with the PDB archive over the course of the coming years and thus provide a unified conduit from experimental models/data to computed models.

Finally, the funding activities since the last renewal were reviewed. The committee notes the huge effort that has gone into raising additional federal grant monies to make up the difference between the funding requested and the funding received at the last renewal. While these efforts were necessary, we do wonder if the funding agencies are best served by requiring this very successful and important community resource to devote so much time and effort to these additional fundraising activities. Maybe a more holistic view of the RCSB funding by the agencies in the upcoming renewal process would serve them and the RCSB better.

Recommendations

- *Provide an update on the outcome of the upcoming NSF/NIH/DOE mid-funding cycle review site visit before the next regular committee meeting.*

Discussion

There was a discussion about what the team should focus on in preparation for the upcoming renewal proposal, to be submitted around February 2023 (actual date TBD, but reasonably extrapolated on the timing of the last renewal). The committee feels that the RCSB is well placed to be successful with a renewal proposal, especially given the stable technical platform, great advances in the structure deposition system, and the increases in biocurator efficiency. The increasing number of cryo-EM derived models presents the RCSB team with a need for increased deposition/curation capacity and data delivery systems to better handle more complex models. This emphasizes the importance of a strong, stable and expanded team.

Recommendations

- *Engage the Advisory Committee (AC) in reviewing the renewal plans early on. Specifically, a virtual meeting should be held with the committee 4-6 months prior to the renewal submission.*

Next Advisory Committee Meeting: scheduling, location

The committee supports the plan for an in-person meeting at Rutgers in Piscataway, NJ, at a time that works for the largest number of committee members and avoids large conferences if possible.

Appendix 1: Responses to 2021 RCSB PDB AC Recommendations

RCSB PDB thanks the Advisory Committee for their participation and thoughtful meeting report.

Our responses to recommendations bulleted in the report follow.

Response to COVID-19

Beyond the recent publication in PROTEINS about the structural biology of the SARS-CoV-2 virus, the RCSB should consider whether there are other venues to highlight the role of the structure, and the PDB, in the pandemic response. Longer term there may be opportunities to study the long-term impact that structural biology and the availability of structures from the PDB has had on the fight against the virus.

RCSB PDB Response: Additional manuscripts related to COVID-19 for publication in peer-reviewed journals have been submitted or are in progress. SARS-CoV-2 structures are also being used as examples in our presentations and in publications about the website. We will also continue to develop related content for PDB-101.

Deposition/Biocuration Recommendations

- *AC: The improvements in deposition efficiency appear to have come in part from the flexibility afforded staff with remote work. The RCSB PDB is encouraged to think about how to continue this model in the future.*
    - RCSB PDB: As our campuses return to “normal”, we will explore how hybrid onsite/offsite model could work best for our team.
- *AC: The committee welcomes the new validation features for cryo-EM depositions. However, the RCSB PDB (in collaboration with the broader wwPDB community) is encouraged to implement new accepted metrics in a timely manner as they become available.*
    - RCSB PDB: We recognize the need to further improve validation of cryo-EM depositions. OneDep team will implement community recommended metrics and software once the PDB has received recommendations from the community.
- *AC: Continue to monitor the growth of cryo-EM depositions and be prepared to prioritize the implementation of deposition standards and tools to help respond to the increased load.*
    - RCSB PDB: In response to the rapid growth of cryo-EM depositions in number and size, 2021-2022 OneDep project planning will be focused on improvements in EM deposition, including better collection of auxiliary data, more controlled vocabularies, more mandatory data items, and better metadata checking.

PDB Archive Status

- *AC: There are several other opportunities to improve the archive through remediation campaigns. The committee encourages the RCSB to prioritize targets such as lipids/detergents, metalloproteins, and assemblies.*
    - RCSB PDB: Delivery of a uniform collection of assembly data files in PDBx/mmCIF format is anticipated in the coming year. Metalloproteins are the next remediation priority on the RCSB PDB roadmap.
- *AC: The remediation of antibody structures is an outstanding problem.*

    - RCSB PDB: The inconsistency in antibody annotation is a recognized limitation in the repository. As a first step in improving this situation, we are incorporating annotations from the Structural Antibody Database (SAbDab) and the International Immunogenetics Information System (IMGT) into RCSB.org to provide an improved delivery and search experience for this molecular class. These annotations should be available in 2021.
- *AC: With the recent technical developments in the field of cryo-electron tomography it seems likely that there will be increased demand for archiving tomography data as part of depositions. The committee wonders whether the current EMDB/EMPIAR system will be sustainable and suggests the RCSB look into how complementary services might be established in the US.*
    - RCSB PDB: RCSB PDB Director will explore this issue with US experts in Cryo-EM/ET.

New and Improved RCSB.org Recommendations

- *AC: Reach out to the Rutgers U/X program to leverage their expertise, and if possible, raise awareness of UX thinking across the team.*
    - RCSB PDB: We have started to discuss options for a collaboration between RCSB PDB and the Rutgers UXD Master's Degree Program that we will target for Spring 2022 that would involve a review of how undergraduate lecturers use RCSB.org tools. This effort will involve a very formal evaluation of RCSB.org UX based on input and data gathered from the undergraduate lecturer cohort and will result in a report containing specific recommendations for site improvements based on the primary use cases for this cohort.
- *AC: Create a 1 year plan for engagement with non-structural biology user communities and updates of user interfaces as a result, including identifying the top 3 communities to target and build relationships with.*
    - RCSB PDB: Through the collaboration with the Rutgers UXD Program we will plan to develop an evaluation process that can reuse and apply with other cohort groups. The prototype group will consist of undergraduate educators, and we plan to extend this to computational and machine learning modelers and life science and clinically focused user groups.
- *AC: Consider the development of a simplified search and visualization interface for the more novice or casual user.*

    - RCSB PDB: We plan to use the engagement of undergraduate educators as part of the Rutgers UXD collaboration as a vehicle to focus on potential strategies to simplify our search and visualization tools. We will also continue to augment documentation, including entry points and examples for current offerings.
- *AC: Consider seeking input from the community as a potentially productive route for thoughts about how to improve the site. Committee members suggested looking at BioMolViz.org and engaging with undergraduate life science education researchers.*
    - RCSB PDB: Through our initial conversations with the Rutgers UXD program, we are currently exploring a project that would involve a review of how undergraduate lecturers use RCSB.org tools. BioMolViz members would be ideal interview subjects for this project.

RCSB PDB: Additional RCSB.org Improvements

A major release in August 2021 included significant changes to the Search User Interface (UI). The main purpose was to enable return of Chemical Components in the search results, with corresponding drilldown options in the Refinement panel. Additional UI enhancements and improvements were bundled with this release. All "search by example" links that appear on RCSB.org open directly in the Query Builder (e.g., from the Structure Summary Page or Annotations tab). This offers users a convenient gateway to the Query Builder refinement options (manual addition of search attributes or Refinement Panel use). Tooltips provide descriptions and examples for each attribute at point-of-use, and where applicable, suggest minimum and maximum input values. This information was previously only available by consulting Help Documentation.

In addition to the specific points identified above, and in response to input that we have received through the Customer Service Help Desk and other sources, we also plan to undertake the following steps to improve the user experience of the new search system in 2022:

- Expand on the "search by example" paradigm, providing more links in strategic places in the website, that give a starting point for both common and more complex queries (e.g., unit cell dimensions).
- Continue providing more common query examples in the existing Search Examples page to serve as a starting point for complex searches. Documentation will continue to be improved and expanded. A "Frequently Run Queries" tab to the Search UI itself (in a tab adjacent to the "History", "Browse Annotations" and "MyPDB" tabs) would allow users to load a pre-constructed

query into the Query Builder without leaving the Search UI. This could include and/or supplement the examples found in the existing Search Examples page

PDB Data Delivery Status

AC: Data Throttling: The consensus was this shouldn't be implemented without community input and analysis of historical data. The committee also felt that the NCBI model for throttling was one that could be adopted if necessary, and that tokens or similar impediments to site access should be avoided.

RCSB PDB Response: Initial steps have been taken to improve the efficiency of our data access and search tools and upgrade supporting infrastructure in order to support reliable and performant services. No matter what capacity building we undertake, there will be situations where a few individual users could overwhelm our capability to deliver a service. For the immediate protection of the majority of our users, we have implemented the NCBI model for throttling.

In the longer term, to meet the needs of valued users with extraordinary service demands we will implement a by invitation only system using API tokens that will better regulate their traffic and avoid impacting overall system availability. We will continue to monitor usage closely as we develop our token-based system anticipated before the end of 2022.

Outreach/Education and PDB50

- *AC: Leverage the PDB50 celebrations for outreach where possible*
    - RCSB PDB: As you have witnessed, the PDB50 celebrations are being used to develop and promote some of our online materials. From May - August, the PDB game, developed by biocurator Brian Hudson, has been accessed 1500 times; the structural biology playing cards by David Goodsell2500 times. The October Molecule of the Month feature will celebrate Fifty Years of Open Access to PDB Structures. Features are included at RCSB.org/pdb50, including links to special article collections from *Journal of Biological Chemistry* and *Nature.*
- *AC: In response to the request for the committee to comment on whether users should be surveyed, and if so, on what topic/s, we saw several potential opportunities. Receiving input on integrated/hybrid methods and their potential users would be valuable in these early stages of the field. It was also felt that obtaining community input on the RCSB.org website features would likely be beneficial. Reaching out to educators for their input is one way to help make sure that the site meets the needs of that important group and by extension many students. The committee suggests gathering focus groups of undergraduate instructors who use the PDB in teaching at national conferences like ASBMB, ASCB, ABRCMS and SACNAS to learn more about how they use the PDB and to even teach them how to use the newer tools that are being developed. We encourage the PDB to host focus groups at national conferences that target specific areas, like COVID-19, cancer research, or drug development.*
    - Through our initial conversations with the Rutgers UXD program, we are currently exploring a project that would involve a review of how undergraduate lecturers use RCSB.org tools. We would like to learn how UXD manages the interview process, and then review to see if and how we could expand this process to other user groups in 2022.

- *AC: Given the likely reduction in the number of conferences in the coming years it would be advantageous to make webinars, with presentations and Q&A sessions a regular RCSB offering.*
    - RCSB PDB: We will develop plans for collecting feedback from users via surveys and focus groups (once conferences are back in person). Initial events include 2 days of webinars with the Royal Society of Chemistry (November 16 and November 18, 2021) on PDB data deposition, validation, and Mol*. Another target will be an online short course being developed with the U.S. National Committee for Crystallography, NIST, and the National Academy of Sciences (planned for April 2022) on the use, development, and maintenance of crystallographic and structural databases. An entire session will be organized by RCSB PDB, wherein feedback will be solicited.
    - These workshops will be supplemented by online surveys of the research community in the second half of 2022.

Operations and Funding Recommendations

- ***Identify a replacement for Andy Byrd from the NMR community in a timely manner***

RCSB PDB Response: We are pleased to announce that Kevin Gardner (CUNY ASRC) will join the committee in Spring 2022 as a replacement for Andy Byrd.

In addition, Lance Stewart of the Institute for Protein Design at UW Medicine has joined the committee, effective Fall 2021.

Discussion: What objectives do we need to reach in preparation for renewal?

- ***Prioritize any groundwork that can be done to make a strong case for funding to support the increase in data volume arising from the growth of techniques like cryo-EM, cryo-ET and hybrid methods.***
- ***Engage the Advisory Committee in reviewing the renewal plans early on***
- ***Where helpful, engage the AC in providing input on the succession planning. At the same time develop plan B so that the renewal proposal isn't impacted if the successor is not in place in time***

RCSB PDB Response: We are pleased that the committee is able to quickly convene in September to hear our revised plans for Q4/2021 and Q1-4/2022.

We look forward to discussing preparations for the next funding period at the Spring meeting.

Next Advisory Committee Meeting

We will work with our funding representatives to find a time and location for Spring 2022.

Appendix 2: PDB 2021 Metrics

In aggregate, 14,571 depositions were received and processed between January 1st and December 31st, 2021, with an average turnaround of two weeks by the wwPDB. This represents a decrease from the 15,436 entries deposited in 2020.

Breakdown of depositions by discipline was as follows:

X-ray:	9,937 (68% of entries deposited, down from 12242 in 2020)
NMR:	364 (2%, down from 390)
EM:	4254 (18%, up from 2,780)
Other:	16 (.1%, down from 26)

Breakdown of depositions by wwPDB processing site was as follows:

RCSB PDB:	5687 (39%)
PDBj:	4160 (29%)
PDBe-EBI:	4724 (32%)

Breakdown of depositors by location was as follows:

North America	34.7%
Europe	32.9%
Asia	27.9%
South America	1.0%
Oceania	3.3%
Africa	<1%

During 2021, RCSB PDB's website at http://rcsb.org was visited by millions of unique visitors. Google Analytics reported ~4.7 million unique visitors accessed RCSB.org in 11 million sessions. Internal statistics tracking platform reports 617.26 TB of Bandwidth.

In 2021, data files from the PDB archive were accessed 2.3 billion times across all wwPDB partner sites.