RCSB Protein Data Bank Advisory Committee

Report of April 21, 2020 Annual Meeting

Teleconference

Chair: Paul Adams

Membership:

Present: Peter Andolfatto, Judith Blake, Andy Byrd, Bridget Carragher, Wah Chiu, Kirk Clark, Paul Craig, Roland Dunbrack, Paul Falkowski, Thomas Ferrin, Mandë Holford, Cathy Peishoff, Sue Rhee, Torsten Schwede

Absent: Robert B. Darnell, Jill Trewhella

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, jill.trewhella@sydney.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 virtual meeting on April 21st, 2020 to review recent progress and provide feedback on specific questions.

Agenda items included

- Welcome and Introductions
- Response to COVID-19
- 2019 Overview
- Deposition/Biocuration
- New and Improved RCSB.org
- Outreach/Education
- Operations and Funding
- PDB50 and AC Meetings
- Discussion: PDB Format

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 (UCSD). Appendix 1 provides a summary of the RCSB responses to the 2019 Advisory Panel meeting recommendations. Appendix 2 provides a summary of global PDB deposition and data access statistics in 2019.

Overall Comments from the Advisory Panel

The team is congratulated on an excellent set of focused presentations, and their collective leadership of the RCSB and their individual projects over the last year. The committee applauds the continued efforts of the RCSB Director, Stephen Burley, and his colleagues to seek additional funding for the RCSB activities. The committee also recognizes that everyone is having to respond to an unprecedented global situation in the form of the COVID-19 pandemic.

Recommendations for future meetings

- Keep up the great work!
- A brief report in the next 2 months on the impact of the COVID-19 pandemic on the RCSB operations and PDB depositions would be very helpful for the committee.

Detailed Advisory Panel Comments and Feedback

COVID-19: Are there other projects in this area we could develop to support research and education?

The current COVID-19 pandemic presents many challenges but also some opportunities for the RCSB. There is a significant interest from researchers and also the general public to learn more about the SARS-CoV-2 virus and pandemics in general. Given the important role that structural biology is playing in understanding the virus, and the role it plays in developing therapeutics, there is a great opportunity for the RCSB to provide material to the community. They have already created the COVID-19/SARS-CoV-2 Resources website (http://rcsb.org/covid19), with links to relevant structures, a general coronavirus education page, educational videos, and many images. The committee also was pleased to hear about plans to host a summer boot camp around the topic of coronavirus in 3D, and the evolution of SARS-CoV-2 proteins. Beyond their COVID-19 current efforts, the RCSB should consider additional activities. The committee identified a number of opportunities, including:

- Creating PDB101s on viral infection processes, immunity, virus-mediated acute respiratory syndrome.
- Extending the current site to provide structural information and links to other material (experiments, recent news, etc) about each protein from the viral genome.
- Enabling or more directly supporting the collection of revisions of structures from the community, which could eventually lead to new version uploads by the original authors.
- Reaching out to the local community to provide information about the basic structure of COVID proteins, viral RNAs, interacting cellular proteins, virus and pathogen in relationship to human diseases through TV news stations, school districts or public health departments.

The committee also recognizes that there may be opportunities to combine education and outreach activities around COVID-19 with fund raising activities, especially with the PDB50 celebrations next year.

Recommendations

- Develop an action plan for expanding the RCSB role in educating the community about COVID-19 and other related pandemics, and the role of structural biology. These plans would ideally be integrated with current and future fundraising activities.
- Track access to the RCSB maintained COVID-19 materials; this would be very helpful for future efforts to highlight the impact of the resource.

Deposition/Biocuration: Any concerns about the Deposition/Biocuration work underway with our wwPDB partners?

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. The load balancing across the wwPDB locations appears to be working well, and new features such as GroupDep are making it easier for some groups to make their depositions. The introduction of mandatory mmCIF deposition for crystallographic models is also hopefully improving the workload on the annotators. A number of other activities across the wwPDB are also likely to improve deposition and biocuration activities in the future, including improved ligand validation and biological assembly annotations, author initiated coordinate replacement, EM map validation, and the remediation of the carbohydrates.

It is also clear that the wwPDB will need to accommodate significant growth in the deposition of atomic resolution models from cryo-EM in the next 5 years. At the same time new XFEL approaches are gaining in popularity. It also currently looks unlikely that there will be dramatic reduction in the number of crystallographic structures deposited each year. This increased volume of structures will need to be processed without a backlog developing. The committee was very pleased to see that an analysis had been performed to provide projections of depositions from 2020 to 2024 for all of the experimental techniques. However, there was a concern that the projections for cryo-EM might be underestimated and not reflective of the current exponential growth.

Recommendations

- 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.
- Continue to track the time taken for depositions, to both measure the load on annotators, and to provide metrics about how process improvements are increasing deposition throughput. This information will be helpful for funding justifications in the future.

New and Improved RCSB.org: Additional Site and Search Functionality requests?

The committee heard from John Westbrook on the ongoing efforts to improve the infrastructure for the rcsb.org website and associated backends. We were impressed how quickly this has been implemented without any substantial interruptions to providing services. Demonstrations of the new search functionality highlighted useful new features. However, there were some concerns about the complexity of the search system for many users, and the loss of important features (such as refining a search to provide a non-redundant set of results). One suggestion was the creation of question-driven functionalities and workflows for popular search activities. The committee recognizes that the RCSB has undertaken community outreach to get user feedback, but these efforts might need to be extended. The new Mol* 3D visualization system was also presented, and clearly shows great potential for interactive display of molecules and maps. However, the committee feels that further development, and in some cases simplification, of the interface would benefit many of the RCSB users.

Recommendations

- Seek further community input, maybe through the creation of focus groups or targeted outreach, to refine the search functionality and the Mol* visualization services.

Outreach/Education: Suggestions for new materials and virtual venues for celebrating PDB50 throughout 2021?

Christine Zardecki presented many of the great efforts in the area of outreach and education over the last year. The PDB-101 resource currently has two-thirds of a million users, and over 2 million page views per year. Significant effort has been spent on developing new content, in particular different aspects of

human health. The committee also was very encouraged to see the results of a community survey, and that 3D print files have been made available for several molecule of the month topics. The community outreach by the team is excellent and clearly very important for educating the research community about the RCSB and structural biology. Of particular importance are the efforts to educate other educators and students. Clearly, the current pandemic will have an impact on the team's ability to perform outreach in person and will require them to develop new approaches to communications. The committee had several suggestions for ways to engage the community in the current circumstances and leverage the importance of structure in the COVID-19 response. One idea was the creation of virtual reality resources, perhaps using ChimeraX. This might provide a platform to propose something similar to Folding@Home where people would be in VR or AR space and attempting to design drugs for COVID-19 proteins. Another suggestion was a competition for creating protein structures from found objects around the home.

The PDB50 celebration in 2021 provides a great opportunity to promote the RCSB widely and emphasize the impact of structural biology. The committee suggested addressing this in multiple forums. Large conferences provide an opportunity for outreach, and in some cases these may be well organized as virtual conferences - the Intelligent Systems for Molecular Biology in 2021 was one example. Museums and other public facing organizations may also provide a great opportunity for engaging a broader audience. The American Museum of Natural History was put forward as an organization looking for online content. Ultimately, the committee feels that there is an opportunity for either local or national recognition through mainstream media, such as NPR and network TV channels. Science Friday at NPR would be a great target for a PDB50 piece, as would a NOVA documentary.

Recommendations

- Create a plan for online outreach and communications for the next 12 months, which incorporates some PDB50 celebration activities
- Develop a PDB50 media communications strategy targeted both locally and nationally.

Next Advisory Committee Meeting location?

While the committee looks forward to the next meeting in person, it seems unlikely that it will come to pass in the first half of 2021. We expect that the next committee meeting will be held virtually. If an in person meeting is possible, the option of coordinating with the PDB50 event at the ACA in Baltimore seems reasonable.

Recommendations

- Plan for a virtual meeting in the first half of 2021.

Discussion: What advice do you have regarding PDB Legacy Format sunsetting?

The committee engaged in a discussion with the RCSB staff on the topic of sunsetting the PDB legacy format, in favor of the mmCIF/PDBx format. The continued support required to serve legacy format files presents an additional burden for the RCSB, and the wwPDB more broadly. **The committee agrees that a timeline for a permanent transition needs to be established, communicated and enacted**. The details of that timeline will require some careful consideration. Firstly, depositions for all structure types will need to move to mmCIF/PDBx before the legacy format can be dropped. Currently, there are plans for cryo-EM submissions to move to mmCIF/PDBx later this year. The committee is still uncertain as to the plans for NMR-based structure depositions. Provided that the milestone of technique-wide mmCIF/PDBx deposition is either reached or scheduled, it seems reasonable to set a firm deadline for dropping support for serving the legacy format files to the RCSB user community. Clearly, extensive outreach and communication should be performed prior to this.

A general transitional approach to the legacy format was also raised by the committee, where the legacy format versions of structures currently available are frozen, and they are not changed any further. This would send the message that mmCIF is the only format capable of handling complex structures but wouldn't remove access to all structures on a specific date, which could present problems for some. This approach also minimizes the work necessary for the RCSB to that of creating a static repository of legacy format structures. The committee did see that there are some additional remaining issues. For example, biological assemblies are currently not available in mmCIF format from RCSB, although they are available from PDBe. If the legacy format is going to be dropped, these files need to be provided sooner rather than later. The mmCIF format is much better than the legacy format for assemblies, which use the MODEL-ENDMDL format to show multiple copies of the ASU in larger assemblies. There is also a good case for the RCSB to generate additional material to help educators know how to best use mmCIF format files. The PDB 101 has a site about beginners using mmCIF/PDBx, which is very helpful. However, it would be even better to have one or two videos for educators that explain how to teach with structures in this file format.

Recommendations

- Develop a timeline for sunsetting the PDB format, which can be widely socialized for feedback. The timeline will need to be consistent with the termination of legacy format model submission for all experimental methods. Consideration could be given to providing an unsupported set of legacy version files to ease the transition for some researchers.
- Make it a priority to address cases where mmCIF files are currently not available, such as assemblies.

Response to the 2020 RCSB PDB Advisory Committee Report
September 17, 2020

RCSB PDB thanks the Advisory Committee for their participation and thoughtful meeting report. Our responses to recommendations bulleted in the report follow.

Executive Summary

- *A brief report in the next 2 months on the impact of the COVID-19 pandemic on the RCSB operations and PDB depositions would be very helpful for the committee.*

RCSB PDB Response:

The majority of RCSB PDB staff have been working remotely since March with no discernible impact on ongoing Operations. All meetings take place regularly via Zoom. The Summer 2020 RCSB PDB Newsletter "Message from the RCSB PDB" article described on-going activities during the pandemic. The team has worked hard and worked very well together despite the challenges of remote working, child care issues, home schooling, *etc*.

The biggest change from our usual sequence of events annually has been on undergraduate research support. Originally, two students were scheduled to perform research on-campus with the RCSB PDB at Rutgers. When programming moved online, we quickly planned a one-week Boot Camp that hosted 31 students followed by a five-week virtual research experience for 12. The results of this research project are currently in preparation for submission to a journal.

As of September 16, 390 SARS-CoV-2 structures have been released in the PDB archive. Each entry has been quickly reviewed and annotated by wwPDB biocurators following these "guiding principles":

- Biocuration of COVID-19 structures is prioritized over that of other structures, including post-release revisions such as citation updates
- Authors are encouraged to release their structures immediately
- Consistent taxonomy name and ID (Severe acute respiratory syndrome coronavirus 2; 2697049) are applied to all COVID-19 structures
- Consistent UniProt referencing is incorporated: P0DTD1, P0DTC1, P0DTC2, P0DTC9

All released SARS-CoV-2 structures and related resources are highlighted at http://RCSB.org/covid19.

Detailed Advisory Panel Comments and Feedback

AC: COVID-19: Are there other projects in this area we could develop to support research and education?

Beyond their COVID-19 current efforts, the RCSB should consider additional activities. The committee identified a number of opportunities, including:

- Creating PDB101s on viral infection processes, immunity, virus-mediated acute respiratory syndrome.
- Extending the current site to provide structural information and links to other material (experiments, recent news, etc) about each protein from the viral genome.
  Enabling or more directly supporting the collection of revisions of structures from the community, which could eventually lead to new version uploads by the original authors.
- Reaching out to the local community to provide information about the basic structure of COVID proteins, viral RNAs, interacting cellular proteins, virus and pathogen in relationship to human diseases through TV news stations, school districts or public health departments.

The committee also recognizes that there may be opportunities to combine education and outreach activities around COVID-19 with fund raising activities, especially with the PDB50 celebrations next year.

Recommendations

- Develop an action plan for expanding the RCSB role in educating the community about COVID-19 and other related pandemics, and the role of structural biology. These plans would ideally be integrated with current and future fundraising activities.
- Track access to the RCSB maintained COVID-19 materials; this would be very helpful for future efforts to highlight the impact of the resource.

RCSB PDB Response:

Since the AC meeting, new SARS-CoV-2 materials at PDB-101 have included

- Molecule of the Month: SARS-CoV-2 Spike and RNA-dependent RNA Polymerase
  New series: Resources to Fight the COVID-19 Pandemic
- Coronavirus Life Cycle painting
- Coronavirus Background for virtual meetings
- Curricula: COVID-19 in Molecular Detail and COVID-19 Evolution and Structural Biology
-

Educational materials are added to the Coronavirus Browse feature at PDB-101. We plan to continue to develop resources throughout the pandemic.

As of June 30, coronavirus-related content accounted for 13% of 1.3 million page views at PDB-101, including

- MOTM main protease: 69,239 views
- MOTM Spike protein: 11,817 views
- Coronavirus images, video, etc.: 99,247 views

In addition, the coronavirus hand-washing video has been viewed >400K times directly on YouTube.

At RCSB.org, the URL http://rcsb.org/covid19 links to all SARS-CoV-2 structures and related resources; this page has been accessed >75K since March 25. Review of RCSB.org traffic shows frequent activity on SARS-CoV-2 pages. The Structure Summary page for the first structure (6lu7) has been accessed >100K times.

We will continue to monitor coronavirus traffic to PDB-101 and RCSB.org, and plan to use the coronavirus story to develop materials for fundraising.

Related SARS-CoV-2 activities have included

- Image contest held in May
- Virtual Boot camp focused on the SARS-CoV-2 Nsp5 main protease held June 22-26 (31 students)
    - Boot camp described in "COVID-19 Evolution and Structural Biology" (2020) *BAMBed,* doi: 10.1002/bmb.21428.
- Undergraduate research experience exploring the full virus June 29-July 30 (12 students)
    - 4 students presented posters at the American Crystallographic Association Meeting
    - 1 student won the MiTeGen-Society of Physics Students Undergraduate Poster Prize

AC: Deposition/Biocuration: Any concerns about the Deposition/Biocuration work underway with our wwPDB partners?

It is also clear that the wwPDB will need to accommodate significant growth in the deposition of atomic resolution models from cryo-EM in the next 5 years. At the same time new XFEL approaches are gaining in popularity. It also currently looks unlikely that there will be dramatic reduction in the number of crystallographic structures deposited each year. This increased volume of structures will need to be processed without a backlog developing. The committee was very pleased to see that an analysis had been performed to provide projections of depositions from 2020 to 2024 for all of the experimental techniques. However, there was a concern that the projections for cryo-EM might be underestimated and not reflective of the current exponential growth.

Recommendations

- 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.

- Continue to track the time taken for depositions, to both measure the load on annotators, and to provide metrics about how process improvements are increasing deposition throughput. This information will be helpful for funding justifications in the future.

RCSB PDB Response:

We shall continue to monitor the growth in cryo-EM depositions closely with the goal of improving our ability to forecast. In addition, we shall continue to evaluate the performance of the wwPDB OneDep System and the efficiency of wwPDB Biocurators for processing cryo-EM depositions and make improvements to the software and our standard operating procedures where indicated.

New and Improved RCSB.org: Additional Site and Search Functionality requests?

The committee heard from John Westbrook on the ongoing efforts to improve the infrastructure for the rcsb.org website and associated backends. We were impressed how quickly this has been implemented without any substantial interruptions to providing services. Demonstrations of the new search functionality highlighted useful new features. However, there were some concerns about the complexity of the search system for many users, and the loss of important features (such as refining a search to provide a non- redundant set of results). One suggestion was the creation of question-driven functionalities and workflows for popular search activities. The committee recognizes that the RCSB has undertaken community outreach to get user feedback, but these efforts might need to be extended. The new Mol* 3D visualization system was also presented, and clearly shows great potential for interactive display of molecules and maps. However, the committee feels that further development, and in some cases simplification, of the interface would benefit many of the RCSB users.

Recommendations
- Seek further community input, maybe through the creation of focus groups or targeted outreach, to refine the search functionality and the Mol* visualization services.

RCSB PDB Response:

Mol*: Since the AC meeting, the user interface has been greatly improved based upon feedback from the community and user documentation was published at RCSB.org. Mol* was used extensively during the summer with students and their feedback was positive. After boot camp, which included many students new to molecular visualization, 65% said Mol* was "easy to use."

RCSB.org: Since April 2020, we have released many additional features (some new and some improved versions of previous offerings), such as returning non-redundant search results, to our searching and reporting services. Due to the nature of the database/software architecture, we were unable to release all services at the same time.

We have started a project to improve documentation to help support users and plan to initiate a user survey at the end of 2020. We are monitoring access to the website to see how it is being used, and are developing tools to better analyze which searches are being performed. These metrics will guide further enhancements of usability/documentation and help prioritize development and release of new features.

AC: Outreach/Education: Suggestions for new materials and virtual venues for celebrating PDB50 throughout 2021?

The committee had several suggestions for ways to engage the community in the current circumstances and leverage the importance of structure in the COVID-19 response. One idea was the creation of virtual reality resources, perhaps using ChimeraX. This might provide a platform to propose something similar to Folding@Home where people would be in VR or AR space and attempting to design drugs for COVID-19 proteins. Another suggestion was a competition for creating protein structures from found objects around the home.

The PDB50 celebration in 2021 provides a great opportunity to promote the RCSB widely and emphasize the impact of structural biology. The committee suggested addressing this in multiple forums. Large conferences provide an opportunity for outreach, and in some cases these may be well organized as virtual conferences - the Intelligent Systems for Molecular Biology in 2021 was one example. Museums and other public facing organizations may also provide a great opportunity for engaging a broader audience. The American Museum of Natural History was put forward as an organization looking for online content. Ultimately, the committee feels that there is an opportunity for either local or national recognition through mainstream media, such as NPR and network TV channels. Science Friday at NPR would be a great target for a PDB50 piece, as would a NOVA documentary.

Recommendations

- Create a plan for online outreach and communications for the next 12 months, which incorporates some PDB50 celebration activities
- Develop a PDB50 media communications strategy targeted both locally and nationally.

RCSB PDB Response:

Select PDB50 activities and materials are being developed in collaboration with our wwPDB partners as listed at https://foundation.wwpdb.org/pdb50.html. The wwPDB is also collaborating on a themed calendar for 2021. Other projects are also being discussed.

In addition, we are developing pitch materials to provide to committee members who have volunteered to help us target museums and media about the PDB using the coronavirus story.

AC: Next Advisory Committee Meeting location?

While the committee looks forward to the next meeting in person, it seems unlikely that it will come to pass in the first half of 2021. We expect that the next committee meeting will be held virtually. If an in person meeting is possible, the option of coordinating with the PDB50 event at the ACA in Baltimore seems reasonable.

Recommendations

- Plan for a virtual meeting in the first half of 2021. The duration of the meeting will be increased to 4 hours (per AC Chair feedback).

RCSB PDB Response:

A doodle poll for available dates in April 2021 will be circulated. Given the AC Chair's other duties, two meeting dates will be confirmed during that month to provide a reserve date in the event that the Chair becomes unavailable on short notice.

AC: Discussion: What advice do you have regarding PDB Legacy Format sunsetting?

The committee agrees that a timeline for a permanent transition needs to be established, communicated and enacted. The details of that timeline will require some careful consideration. Firstly, depositions for all structure types will need to move to mmCIF/PDBx before the legacy format can be dropped. Currently, there are plans for cryo-EM submissions to move to mmCIF/PDBx later this year. The committee is still uncertain as to the plans for NMR-based structure depositions. Provided that the milestone of technique-wide mmCIF/PDBx deposition is either reached or scheduled, it seems reasonable to set a firm deadline for dropping support for serving the legacy format files to the RCSB user community. Clearly, extensive outreach and communication should be performed prior to this.

A general transitional approach to the legacy format was also raised by the committee, where the legacy format versions of structures currently available are frozen, and they are not

changed any further. This would send the message that mmCIF is the only format capable of handling complex structures but wouldn't remove access to all structures on a specific date, which could present problems for some. This approach also minimizes the work necessary for the RCSB to that of creating a static repository of legacy format structures. The committee did see that there are some additional remaining issues. For example, biological assemblies are currently not available in mmCIF format from RCSB, although they are available from PDBe. If the legacy format is going to be dropped, these files need to be provided sooner rather than later. The mmCIF format is much better than the legacy format for assemblies, which use the MODEL-ENDMDL format to show multiple copies of the ASU in larger assemblies. There is also a good case for the RCSB to generate additional material to help educators know how to best use mmCIF format files. The PDB 101 has a site about beginners using mmCIF/PDBx, which is very helpful. However, it would be even better to have one or two videos for educators that explain how to teach with structures in this file format.

Recommendations

- Develop a timeline for sunsetting the PDB format, which can be widely socialized for feedback. The timeline will need to be consistent with the termination of legacy format model submission for all experimental methods. Consideration could be given to providing an unsupported set of legacy version files to ease the transition for some researchers.

- Make it a priority to address cases where mmCIF files are currently not available, such as assemblies.

RCSB PDB Response:

At the wwPDB AC Meeting in October 2020, the wwPDB PIs will ask the advisors to concur with work with the user community to understand how many rely on "best efforts" PDB Legacy format files and how they use these files in their day-to-day research and teaching activities.

The wwPDB will analyze the results of the survey and develop plans for deprecation of the legacy PDB file format that minimizes disruption to users and promotes the FAIR principles of findability, accessibility, interoperability, and reusability. The wwPDB will also determine how best to help our diverse user community transition to the PDBx/mmCIF format regime with webinars, conversion software, on-line tutorials, etc.