Protein Data Bank (PDB) Archive and the wwPDB Partnership

Stephen K. Burley, M.D., D.Phil.
Worldwide Protein Data Bank Co-Leader
Director, RCSB Protein Data Bank
Rutgers University/UC San Diego
Protein Data Bank

- Global archive of experimental macromolecular structure data central to basic and applied research and education in biology and biomedicine
- First open access digital data resource in biology (est. 1971 with 7 entries)
- Single global archive of experimental 3D structures of proteins, DNA, and RNA (>124,000 entries today)
  - Primary data for structural biology, computational biology, drug discovery, …
  - Complements GenBank and UniProt sequence databases
- All data freely available without restrictions on usage (we adhere to the FAIR Principles)

ABL tyrosine-kinase inhibited by Imatinib for treatment of chronic myeloid leukemia (CML).

HIV-1 reverse transcriptase complex with DNA and nevirapine
wwPDB Partners/Responsibilities

- Partners share “Data In” responsibilities
  - Biocurate new depositions
  - Define deposition and annotation policies
  - Implement community validation standards
- Partners distribute identical data via FTP
- Partners provide complementary “Data Out” resources
- Advised by an international committee of experts
PDB Facts and Figures

- Archival Contents
  - >124,000 Structures Released since 1971
  - ~11,000 New Structures Deposited/Year

- Global User Base
  - ~30,000 Depositors Worldwide
  - >1 Million Unique Users/Year
    from 192/195 UN-recognized sovereign nations

- Impacts all of Biology and Medicine
  - >500 Million Data Files Downloaded/Year
  - ~1.5 Million Data Files Downloaded/Day
  - >200 derived data resources repackage PDB data
1. wwPDB Operations/Funding

Founding wwwPDB Partners operate collaboratively from regional data centers in US, EU, and Asia

- US-based RCSB PDB is funded from US sources
- EU-based PDBe is funded from UK and EU sources
- Japan-based PDBj is funded from Japanese source

BioMagResBank (BMRB) joined wwPDB in 2006

- BMRB US operations funded from US sources
- BMRB Japan operations funded from Japanese source
1. wwPDB Funding Sources

- **RCSB PDB (USA):** Core Operations – NSF-administered Cooperative Agreement (~90%; 5yrs; NSF/NIH/DoE); Value-added Activities – NIH/NSF (~10%; 3-5yrs)

- **PDBe (Europe):** Core Operations – EBML (~50%; 3-5yrs), BBSRC (~6%; 1-2yrs), and Wellcome Trust (~20%; 5yrs); Value-added Activities – EMBL/EU/BBSRC/Wellcome Trust (~24%; 1-3yrs)

- **PDBj (Asia):** Core Operations – JST NBDC (~80%; 3yrs) Value-added Activities – MEXT (~20%; 1-3yrs)

- **BMRB_{Madison}:** Core Operations – NIH NIGMS (~80%; 5yrs) Value-added Activities – NIH NIGMS (~20%; 5yrs)

- **BMRB_{Osaka}:** Core Operations – JST NBDC (~80%; 3yrs) Value-added Activities – MEXT (~20%; 1-3yrs)
2. User Communities

- **Data Producers:** Structural biologists worldwide (crystallography, nuclear magnetic resonance, and electron microscopy)

- **Primary Data Consumers:** Researchers, Educators, and Students worldwide in basic and applied biology, medicine, allied health professions, bioinformatics, chemistry, physics, engineering, computer science, statistics/biostatistics, materials science, mathematics, plant sciences, animal husbandry, ecology, …

- **Other Data Consumers:** Patients, Patient’s Families, Patient Advocates, Artists, Journalists, Media Outlets, and Curious Public worldwide
2. User Communities

- **Data Producers:**
  ~30,000 registered depositors adding ~11,000 new structures/yr (Archive growth ~9%/yr)

- **Data Consumers:**
  >1 million unique users/yr

- **Data Rates:** ~1.5 million data files downloaded/day; >500 million data files downloaded/year

- **Pharmaceutical Cos**
  use PDB archive inside company firewalls daily
3. Data Utilization/Monitoring

- **Basic and Applied Researchers** in every area of biology, medicine, and biotechnology, and selected areas of chemistry, physics, materials science, engineering, computer science, statistics/biostatistics, and mathematics worldwide

- **Educators and Students**
  - Graduate/Professional Schools
  - Technical/Undergraduate Colleges
  - Schools (Kindergarten → High School)

- **Other Users:** Science Funders, > 200 Derived Data Resources, Patients, Patient’s Families, Patient Advocates, Artists, Journalists, Media Outlets, and the Curious Public worldwide
3. Data Utilization/Monitoring

- Citation of wwPDB Peer Reviewed Publications
  - >20,000 total citations and ~2000 annual citations of publications from wwPDB Partners
  - Track Citations versus Research Areas
- PDB Structure Identifiers now being “cited” nearly as frequently as our peer reviewed publications
- Protein Data Bank appears in >3,000 issued US patents
- PDB Data File Downloads
  - wwPDB FTP Sites: ~370 million in 2015
  - wwPDB Partner Websites: ~165 million in 2015
4. Impact of Losing PDB Archive

- Current PDB holdings exceed 124,000 experimentally determined 3D structures of biological macromolecules.

- Estimated cost of replicating each PDB entry ranges from US$50,000 to >US$250,000.

- Conservative cost of replicating the PDB archive (assuming an average unit cost of US$100,000) gives:

  PDB Replacement Cost > US$12.5 billion.
4. Impact of Losing PDB Archive

- Research progress would be slowed in every area of biology and medicine and related fields worldwide.
- More than 200 Derived Data Resources (e.g., UniProt, model organism databases) would no longer be able to repackage PDB data.
- Drug discovery innovation in the pharmaceutical and biotechnology industry would be slowed (impacting work on novel targets and mechanisms of action, and NCEs).
- Biology and medical education in schools, colleges, research universities, and graduate and profession schools would be compromised.
- Patients, Patient’s Families, and Patient Advocates forced to make less informed choices re treatment/management.
5. Contingency Planning

- **Scenario 1: Single wwPDB Partner loses funding**
  - Surviving wwPDB Partners could continue providing global data access
  - Surviving wwPDB Partners would need additional funding to shoulder increased data deposition/biocuration/validation burdens coming from geographic area no longer served

- **Scenario 2: All wwPDB Partners lose funding**
  - Current PDB archive FTP tree contents could be transferred to a data commons
  - All data deposition/biocuration/validation operations would cease → no growth in the PDB archive
6. Challenges Facing the wwPDB

1. Year-on-Year growth in number of PDB depositions and their complexity (needing more human biocuration effort)

2. Technology/infrastructure required to keep pace with rapid methodological advances (e.g., Hybrid methods, XFEL)

3. Short duration of current funding cycles

4. Current funding mechanisms are not fit for purpose
   - Tailored for 3-5 year duration research grants
   - Emphasis on discovery versus meeting infrastructure needs

5. Declining funding levels in all wwPDB Partner geographies over the past decade plus

6. Lack of distinction in the Big Data Resource Funding Debate between Primary Data Archival Resources and Secondary Data Resources that aggregate other peoples data (i.e., Huffington Post’s of Biology)
6. Challenges Facing the wwPDB

Common locations where data from NIH-funded work published in 2011 was shared, based on PubMed SI field and PMC ACKnowledgements

1. clinicaltrials.gov
2. PDB
3. GEO
4. Genbank
5. Pubchem
6. IRD
7. MGI
8. DIP

Protein Data Bank (PDB) Archive and the wwPDB Partnership

Worldwide Protein Data Bank Co-Leaders
Stephen K. Burley, M.D., D.Phil. — RCSB PDB
  John L. Markley, Ph.D. — BMRB
  Haruki Nakamura, Ph.D. — PDBj
  Sameer S. Velankar, Ph.D. — PDBe

wwpdb.org