Measuring the Impact of the Protein Data Bank

Helen M Berman

February 28, 2017
Protein Data Bank

- First open access digital resource for biology data (est. 1971 with 7 entries)
- Single global archive of experimental 3D structures of biological macromolecules (>126,000 entries)
  - Primary data for structural biology, computational biology, drug discovery, ...
  - Complements GenBank and UniProt sequence database
- All data made freely available (primary users scientists and educators around the globe)
- Global archive of experimental macromolecular structure data central to biomedical research

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

HIV-1 reverse transcriptase complex with DNA and nevirapine
Organizational Structure/Funding

- Partnerns share “Data In” responsibilities
  - Biocurate new depositions
  - Define deposition and annotation policies
  - Resolve data representation issues
  - Implement community validation standards

- Partners independently funded by each region

- Overseen by a wwwPDB Advisory Committee

- Partners compete on “Data Out” resources
PDB Archive Facts and Figures

- **Archival Contents**
  - ~126,000 Structures Released since 1971
  - ~11,000 New Structures Deposited/Year

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

- **Impacts all of Biology and Medicine**
  - >590 Million Data Files Downloaded/Year
  - ~1.6 Million Data Files Downloaded/Day
  - >200 derived data resources repackage PDB data
## Download Statistics

<table>
<thead>
<tr>
<th>Year</th>
<th>Total</th>
<th>Total FTP Archive</th>
<th>Total Website</th>
<th>RCSB PDB FTP Archive</th>
<th>RCSB PDB Website</th>
<th>PDB FTP Archive</th>
<th>PDB Website</th>
<th>PDBj FTP Archive</th>
<th>PDBj Website</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>294,326,976</td>
<td>213,180,966</td>
<td>81,146,010</td>
<td>159,248,214</td>
<td>64,569,658</td>
<td>34,383,219</td>
<td>14,017,349</td>
<td>19,549,533</td>
<td>2,559,003</td>
</tr>
<tr>
<td>2011</td>
<td>383,131,048</td>
<td>276,952,286</td>
<td>106,178,762</td>
<td>204,939,406</td>
<td>81,560,098</td>
<td>40,960,368</td>
<td>18,515,245</td>
<td>31,052,512</td>
<td>6,103,419</td>
</tr>
<tr>
<td>2013</td>
<td>441,262,210</td>
<td>296,176,290</td>
<td>145,085,920</td>
<td>215,331,908</td>
<td>97,549,580</td>
<td>43,684,850</td>
<td>37,762,496</td>
<td>37,159,532</td>
<td>9,773,844</td>
</tr>
<tr>
<td>2015</td>
<td>534,339,871</td>
<td>368,244,766</td>
<td>166,095,105</td>
<td>255,346,630</td>
<td>111,802,897</td>
<td>48,544,330</td>
<td>41,127,219</td>
<td>64,353,806</td>
<td>13,164,989</td>
</tr>
</tbody>
</table>

More than 1.6 million / day

*Geographic origins of FTP downloads, 2012-2015*
Impact: Primary RCSB PDB Publication

The Protein Data Bank
Helen M. Berman1,2,*, John Westbrook1,2, Zuankun Feng1,2, Gary Gilliland1,3, T. N. Bhat1,3, Helge Weissig1,4, Ilya N. Shindyalov1 and Philip E. Bourne1,4,5

1Research Collaboratory for Structural Bioinformatics (RCSB), 2Department of Chemistry, Rutgers University, 610 Taylor Road, Piscataway, NJ 08854-8087, USA, 3National Institute of Standards and Technology, Route 270, Quince Orchard Road, Gaithersburg, MD 20899, USA, 4San Diego Supercomputer Center, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92030-0701, USA, 5Department of Pharmacology, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92030-0700, USA and 6The Burnham Institute, 15911 North Torrey Pines Road, La Jolla, CA 92037, USA

Received September 20, 1999; Revised and Accepted October 17, 1999

Cited by 21459
Cited ~1500 times/year

<table>
<thead>
<tr>
<th>Field: Research Areas</th>
<th>Record Count</th>
<th>% of 15137</th>
<th>Bar Chart</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIOCHEMISTRY MOLECULAR BIOLOGY</td>
<td>7907</td>
<td>52.236%</td>
<td></td>
</tr>
<tr>
<td>CHEMISTRY</td>
<td>3075</td>
<td>20.314%</td>
<td></td>
</tr>
<tr>
<td>BIOPHYSICS</td>
<td>2823</td>
<td>18.650%</td>
<td></td>
</tr>
<tr>
<td>COMPUTER SCIENCE</td>
<td>2310</td>
<td>15.261%</td>
<td></td>
</tr>
<tr>
<td>PHARMACOLOGY PHARMACY</td>
<td>1962</td>
<td>12.962%</td>
<td></td>
</tr>
<tr>
<td>MATHEMATICAL COMPUTATIONAL BIOLOGY</td>
<td>1596</td>
<td>10.544%</td>
<td></td>
</tr>
<tr>
<td>BIOTECHNOLOGY APPLIED MICROBIOLOGY</td>
<td>1258</td>
<td>8.311%</td>
<td></td>
</tr>
<tr>
<td>SCIENCE TECHNOLOGY OTHER TOPICS</td>
<td>810</td>
<td>5.351%</td>
<td></td>
</tr>
<tr>
<td>CRYSTALLOGRAPHY</td>
<td>762</td>
<td>5.034%</td>
<td></td>
</tr>
<tr>
<td>PHYSICS</td>
<td>695</td>
<td>4.591%</td>
<td></td>
</tr>
<tr>
<td>MATHEMATICS</td>
<td>648</td>
<td>4.281%</td>
<td></td>
</tr>
<tr>
<td>CELL BIOLOGY</td>
<td>609</td>
<td>4.023%</td>
<td></td>
</tr>
<tr>
<td>GENETICS HEREDITY</td>
<td>390</td>
<td>2.576%</td>
<td></td>
</tr>
<tr>
<td>ENGINEERING</td>
<td>371</td>
<td>2.451%</td>
<td></td>
</tr>
<tr>
<td>LIFE SCIENCES BIOMEDICINE OTHER TOPICS</td>
<td>240</td>
<td>1.586%</td>
<td></td>
</tr>
<tr>
<td>MICROBIOLOGY</td>
<td>177</td>
<td>1.169%</td>
<td></td>
</tr>
<tr>
<td>IMMUNOLOGY</td>
<td>166</td>
<td>1.097%</td>
<td></td>
</tr>
<tr>
<td>MATERIALS SCIENCE</td>
<td>159</td>
<td>1.050%</td>
<td></td>
</tr>
<tr>
<td>RESEARCH EXPERIMENTAL MEDICINE</td>
<td>120</td>
<td>0.793%</td>
<td></td>
</tr>
<tr>
<td>PLANT SCIENCES</td>
<td>118</td>
<td>0.780%</td>
<td></td>
</tr>
<tr>
<td>SPECTROSCOPY</td>
<td>112</td>
<td>0.740%</td>
<td></td>
</tr>
<tr>
<td>POLYMER SCIENCE</td>
<td>96</td>
<td>0.634%</td>
<td></td>
</tr>
</tbody>
</table>
3166 Patents Mention “protein data bank”

1. 9,476,035 Recombinant polymerases with increased phototolerance
2. 9,475,886 Recombinant antibody composition
3. 9,475,881 Antibody variants with enhanced complement activity
4. 9,475,862 Neutralizing GP41 antibodies and their use
5. 9,475,851 High MAST2-affinity polypeptides and uses thereof
6. 9,475,847 Insecticidal proteins and methods for their use
7. 9,474,759 Broad-spectrum antivirals against 3C or 3C-like proteases of picornavirus-like supercluster: picornaviruses, caliciviruses and coronaviruses
8. 9,469,684 Therapeutic and diagnostic cloned MHC-unrestricted receptor specific for the MUC1 tumor associated antigen
9. 9,468,660 Antinematodal methods and compositions
10. 9,464,311 Method for identifying modulators of ubiquitin ligases
11. 9,464,280 Beta-lactamases with improved properties for therapy
12. 9,458,470 Recombinant influenza virus-like particles (VLPs) produced in transgenic plants expressing hemagglutinin
13. 9,458,434 Mutant enzyme and application thereof
14. 9,458,229 Immunogenic proteins and compositions
15. 9,453,236 Polynucleotides and polypeptides involved in post-transcriptional gene silencing
16. 9,453,224 MiRNA modulators of thermogenesis
17. 9,453,019 Linked purine pterin HPPK inhibitors useful as antibacterial agents
18. 9,452,222 Nucleic acids encoding modified relaxin polypeptides
19. 9,452,210 Influenza virus-like particles (VLPS) comprising hemagglutinin produced within a plant
20. 9,451,783 Phytase variants
21. 9,447,157 Nitration shielding peptides and methods of use thereof
22. 9,447,156 Methods and compositions for inhibiting neddylation of proteins
23. 9,447,127 Synthetic lung surfactant and use thereof
24. 9,446,121 Cloning of honey bee allergen
25. 9,446,116 Peptide sequences and compositions
26. 9,443,017 System and method for displaying search results

http://patft.uspto.gov/
Accessed October 26, 2016
Impact: PDB Data Reuse

- PDB data used by >200 biological databases
  - Based on databases publishing in *NAR* 2011-2016
  - 11 Categories: Structure, Protein Sequence, Nucleotide Sequence, RNA Sequence, Genomics, Metabolic and Signaling, Human Genes and Diseases, Immunology, Proteomics, Plant, Other

- Since 2011, >25% of new databases utilize PDB data (119 out of 452 new databases)
Cost of Replicating the PDB Archive

- Data integrity and security are of paramount importance to the wwPDB partnership
- Estimated cost of replicating each PDB entry ranges from US$50,000 to > US$250,000
- Cost of replicating the PDB archive: **US$12 billion**
  (assuming <unit cost>=US$100,000)
What Has the PDB Archive Enabled?

- Reproducibility and secure storage
- Accelerated structure determination technologies
- Understanding evolution in 3D
  - Structure classification and prediction
- Creation of structural bioinformatics as a discipline
- Structure-based drug discovery
- Functional understanding of biology at molecular and atomic levels

Antigen Presenting Cell meets the T-cell
PDB 2CKB, Garcia et al. (1998)