Measuring the Impact of the Protein Data Bank

Helen M Berman

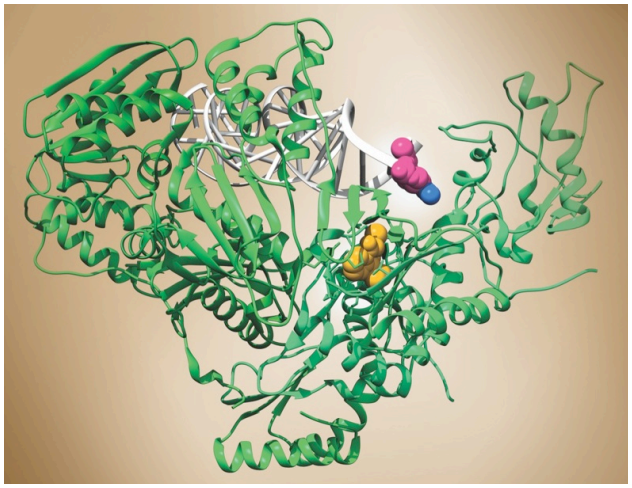
February 28, 2017

WORLDWIDE PDB
PROTEIN DATA BANK

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).
PDB ID 2hyy Cowan-Jacob *et al.* (2007) *Acta Crystallographica D63*: 80-93.



HIV-1 reverse transcriptase complex with DNA and nevirapine
PDB ID 3v81 Das *et al.* (2012) *Nature Structural and Molecular Biol ogy19:* 253-259.

Organizational Structure/Funding

wwpdb.org

RCSB PDB PROTEIN DATA BANK
US: NSF, NIH, DOE

PDBe
Protein Data Bank in Europe
Europe: EMBL-EBI, Wellcome Trust, BBSRC, EU, NIH, CCP4

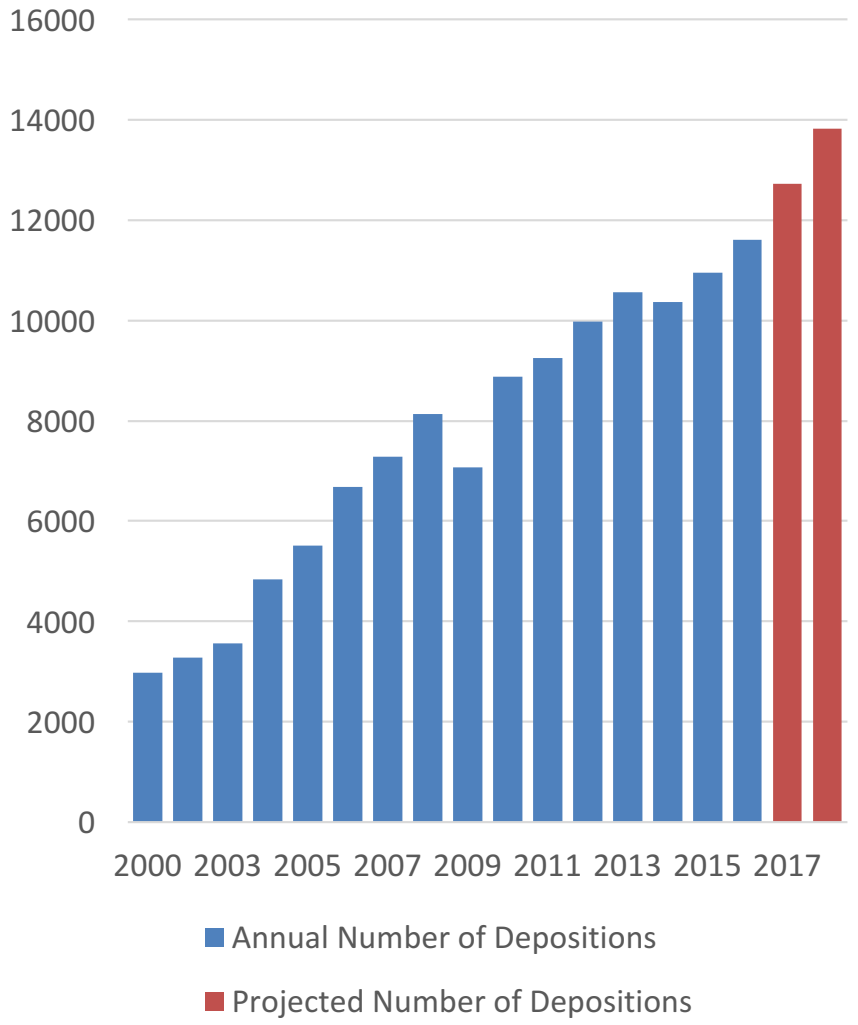
PDBj
Protein Data Bank Japan
Asia: NBDC-JST

BMRB
US: NIGMS

- Partners 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 wwPDB 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

Year	Total	Total FTP Archive	Total Website	RCSB PDB FTP Archive	RCSB PDB Website	PDBe FTP Archive	PDBe Website	PDBj FTP Archive	PDBj Website
2010	294,326,976	213,180,966	81,146,010	159,248,214	64,569,658	34,383,219	14,017,349	19,549,533	2,559,003
2011	383,131,048	276,952,286	106,178,762	204,939,406	81,560,098	40,960,368	18,515,245	31,052,512	6,103,419
2012	376,944,070	255,837,735	121,106,335	213,510,347	90,438,501	21,601,103	23,982,801	20,726,285	6,685,033
2013	441,262,210	296,176,290	145,085,920	215,331,908	97,549,580	43,684,850	37,762,496	37,159,532	9,773,844
2014	512,227,251	339,193,721	173,033,530	237,168,615	110,115,316	52,362,370	48,031,414	49,662,736	14,886,800
2015	534,339,871	368,244,766	166,095,105	255,346,630	111,802,897	48,544,330	41,127,219	64,353,806	13,164,989
2016	**591,876,087**	**366,677,897**	**225,198,190**	**293,648,366**	**161,208,456**	**30,274,284**	**44,432,830**	**42,755,247**	**19,556,904**

More than 1.6 million / day

Geographic origins of FTP downloads, 2012-2015

Impact: Primary RCSB PDB Publication

Nucleic Acids Research, 2000, Vol. 28, No. 1 **235–242**

The Protein Data Bank

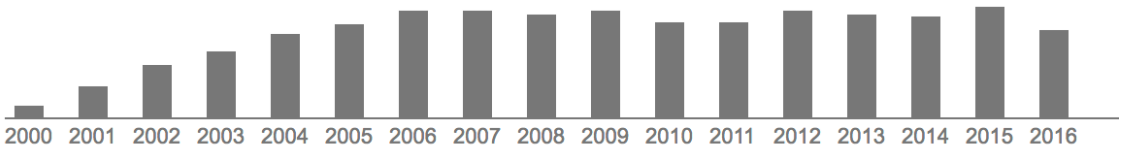
Helen M. Berman[1,2,*], John Westbrook[1,2], Zukang Feng[1,2], Gary Gilliland[1,3], T. N. Bhat[1,3], Helge Weissig[1,4], Ilya N. Shindyalov[4] and Philip E. Bourne[1,4,5,6]

[1]Research Collaboratory for Structural Bioinformatics (RCSB), [2]Department of Chemistry, Rutgers University, 610 Taylor Road, Piscataway, NJ 08854-8087, USA, [3]National Institute of Standards and Technology, Route 270, Quince Orchard Road, Gaithersburg, MD 20899, USA, [4]San Diego Supercomputer Center, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0505, USA, [5]Department of Pharmacology, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0500, USA and [6]The Burnham Institute, 10901 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



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

NEWS FEATURE

THE TOP 100 PAPERS

Nature explores the most-cited research of all time.

BY RICHARD VAN NOORDEN, BRENDAN MAHER AND REGINA NUZZO

The discovery of high-temperature superconductors, the determination of DNA's double-helix structure, the first observations that the expansion of the Universe is accelerating — all of these breakthroughs won Nobel prizes and international acclaim. Yet none of the papers that announced them comes anywhere close to ranking among the 100 most highly cited papers of all time.

Citations, in which one paper refers to earlier works, are the standard means by which authors acknowledge the source of their methods, ideas and findings, and are often used as a rough measure of a paper's importance. Fifty years ago, Eugene Garfield published the Science Citation Index (SCI), the first systematic effort to track citations in the scientific literature. To mark the anniversary, *Nature* asked Thomson Reuters, which now owns the SCI, to list the 100 most highly cited papers of all time. (See the full list at www.nature.com/top100.) The search covered all of Thomson Reuter's Web of Science, an online version of the SCI that also includes databases covering the social sciences, arts and humanities, conference proceedings and some books. It lists papers published from 1900 to the present day.

The exercise revealed some surprises, not least that it takes a staggering 12,119 citations to rank in the top 100 — and that many of the world's most famous papers do not make the cut. A few that do, such as the first observation of carbon nanotubes (number 36) are indeed classic discoveries. But the vast majority describe experimental methods or software that have become essential in their fields.

The most cited work in history, for example, is a 1951 paper[1] describing an assay to determine the amount of protein in a solution. It has now gathered more than 305,000 citations — a recognition that always puzzled its lead author, the late US biochemist Oliver Lowry. "Although I really know it is not a great paper … I secretly get a kick out of the response," he wrote in 1977.

The colossal size of the scholarly literature means that the top-100 papers are extreme outliers. Thomson Reuter's Web of Science holds some 58 million items. If that corpus were scaled to Mount Kilimanjaro, then the 100 most-cited papers would represent just 1 centimetre at the peak. Only 14,499 papers — roughly a metre and a half's worth — have more than 1,000 citations (see 'The paper mountain'). Meanwhile, the foothills comprise works that have been cited only once, if at all — a group that encompasses roughly half of the items.

Nobody fully understands what distinguishes the sliver at the top from papers that are merely very well known — but researchers' customs explain some of it. Paul Wouters, director of the Centre for Science and Technology Studies in Leiden, the Netherlands, says that many methods papers "become a standard reference that one cites in order to make clear to other scientists what kind of work one is doing". Another common practice in science ensures that truly foundational discoverers — Einstein's special theory of relativity, for instance — get fewer citations than they might deserve: they are so important that they quickly enter the textbooks or are incorporated into the main text of papers as terms deemed so familiar that they do not need a citation.

Citation counts are riddled with other confounding factors. The volume of citations has increased, for example — yet older papers have had more time to accrue citations. Biologists tend to cite one another's work more frequently than, say, physicists. And not all fields produce the same number of publications. Modern bibliometricians therefore recoil from methods as crude as simply counting citations when they want to measure a paper's value: instead, they prefer to compare counts for papers of similar age, and in comparable fields.

Nor is Thomson Reuters' list the only ranking system available. Google Scholar compiled its own top-100 list for *Nature*. It is based on many more citations because the search engine culls references from a much greater (although poorly characterized) literature base, including from a large range of books. In that list, available at www.nature.com/top100, economics papers have more prominence. Google Scholar's list also features books, which Thomson Reuters did not analyse. But among the science papers, many of the same titles show up.

Yet even with all the caveats, the old-fashioned hall of fame still has value. If nothing else, it serves as a reminder of the nature of scientific knowledge. To make exciting advances, researchers rely on relatively unsung papers to describe experimental methods, databases and software.

Here *Nature* tours some of the key methods that tens of thousands of citations have hoisted to the top of science's Kilimanjaro — essential, but rarely thrust into the limelight.

BIOLOGICAL TECHNIQUES

For decades, the top-100 list has been dominated by protein biochemistry. The 1951 paper[1] describing the Lowry method for quantifying protein remains practically unreachable at number 1, even though many biochemists say that it and the competing Bradford assay[2] — described by paper number 3 on the list — are a tad outdated. In between, at number 2, is Laemmli buffer[3], which is used in a different kind of protein analysis. The dominance of these techniques is attributable to the high volume of citations in cell and molecular biology, where they remain indispensable tools.

At least two of the biological techniques described by top-100 papers have resulted in Nobel prizes. Number 4 on the list describes the DNA-sequencing method[4] that earned the late Frederick Sanger his share of the 1980 Nobel Prize in Chemistry. Number 63 describes polymerase chain reaction ➤

3166 Patents Mention “protein data bank”

USPTO PATENT FULL-TEXT AND IMAGE DATABASE

Home | Quick | Advanced | Pat Num | Help

Next List | Bottom | View Cart

Searching US Patent Collection...

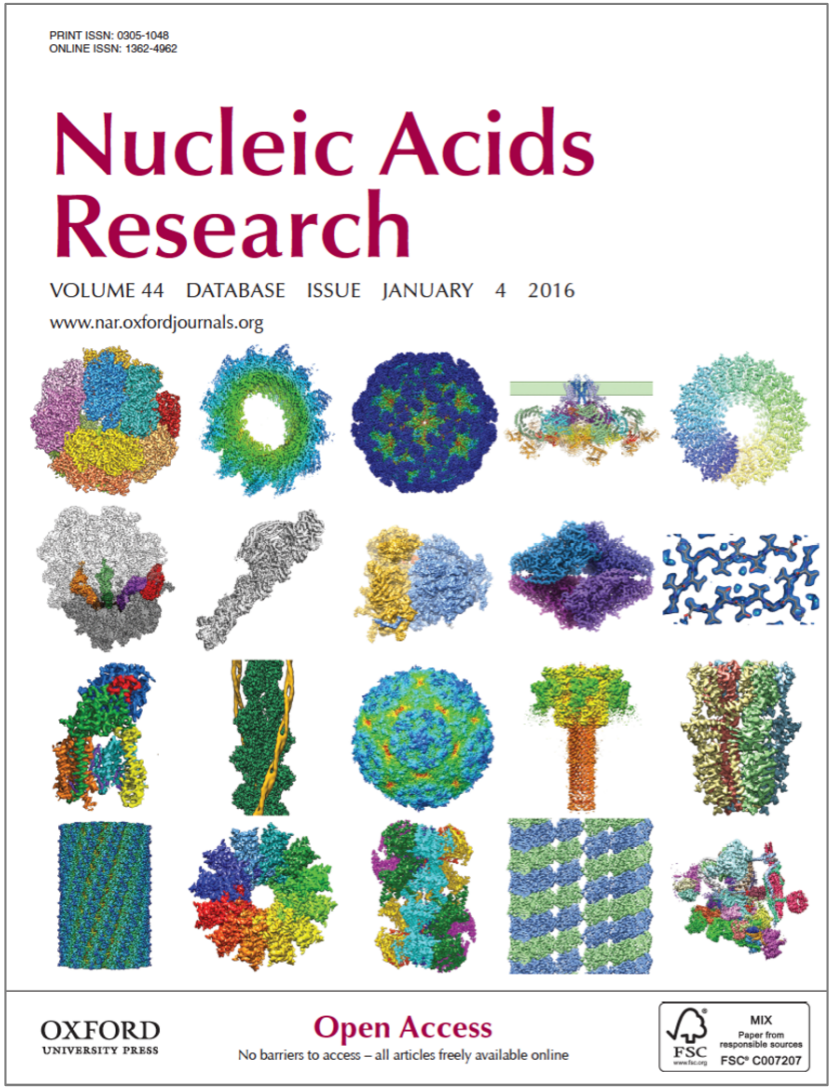
Results of Search in US Patent Collection db for:
"**protein data bank**": 3166 patents.
*Hits **1** through **50** out of **3166***

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)

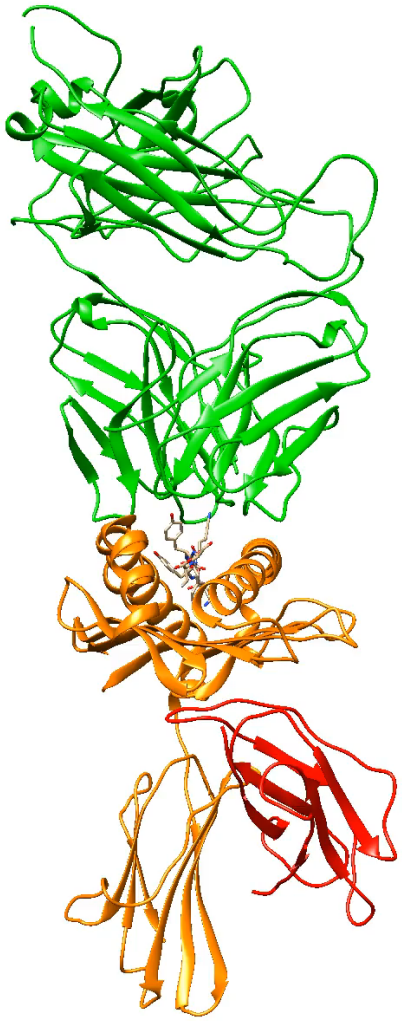


Cost of Replicating the PDB Archive

- 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)