Visualizing the Impact of a Journal Article: "The Protein Data Bank" in *Nucleic Acids Research*, 2000

Susan Makar, M.L.I.S.
Research Librarian
National Institute of Standards and Technology

Amanda Malanowski, B.S. Trad. Math.
Program Analyst
National Institute of Standards and Technology

Talapady Bhat. Ph.D.
Research Chemist
National Institute of Standards and Technology

Introduction

The Protein Data Bank (PDB) was established at Brookhaven National Laboratories in 1971 as an archive for biological macromolecular crystal structures. Originally, it contained only seven structures; today it holds over 129,000 structures of large biological molecules, including proteins and nucleic acids. In 1998, the management of the PDB became the responsibility of the Research Collaboratory for Structural Bioinformatics (RCSB), a consortium composed of Rutgers, the State University of New Jersey; the University of California at San Diego; and the National Institute of Standards and Technology (NIST)[1]. Since 2005, RCSB includes two members, Rutgers and the University of California at San Diego. In 2000, two NIST researchers, Dr. Talapady Bhat and Dr. Gary Gilliland, along with researchers from other institutions, co-authored the article "The Protein Data Bank" in *Nucleic Acids Research* (volume 28, issue 1, pages 235-242). It has since become NIST's most highly cited journal article.

In collaboration with Dr. Bhat, one of the NIST co-authors of "The Protein Data Bank," library staff in the Information Services Office (ISO) at NIST analyzed this article, studying the authors, institutions, journals, research areas, and countries that have cited the article. ISO staff used library resources and tools to analyze the paper and visualize its impacts. ISO is responsible for creating, maintaining, organizing, and disseminating information to support the research and programmatic needs required to fulfill the scientific and technical mission of NIST.

[1] NIST is a non-regulatory federal agency within the U.S. Department of Commerce (DOC). NIST's mission is to promote U.S. innovation and industrial competitiveness by advancing measurement science, standards, and technology in ways that enhance economic security and improve the quality of life.

Methodology

Web of Science[2] (WoS), a database that indexes and abstracts the peer-reviewed scientific and technical literature, was searched to identify papers citing "The Protein Data Bank" since it was published in 2000. This collection of citing papers was then studied using the "Analyze Results" feature in WoS to identify all the authors, countries, institutions, research areas, and source titles (journals) citing "The Protein Data Bank." The results for each of these analyses (authors, countries, etc.) were exported from WoS into *Excel*.

Once the data was exported into *Excel*, the citing papers were analyzed using *Excel* and the data visualization software *Tableau* to create this paper's graphics. *Tableau* was used to generate a bubble map, a tree map, and a global citation map to show the impact of "The Protein Data Bank." *Excel* was used to create a bar graph and a pie chart.

Findings

The analysis, based on a *Web of Science* (WoS) search on March 22, 2017, yielded 15,622 citations to "The Protein Data Bank" since it was published in 2000. A Google Scholar search, which includes the grey literature, yielded 22,377 citations on the same day. The citation pattern in Figure 1 shows that "The Protein Data Bank" has remained relevant since it was published.

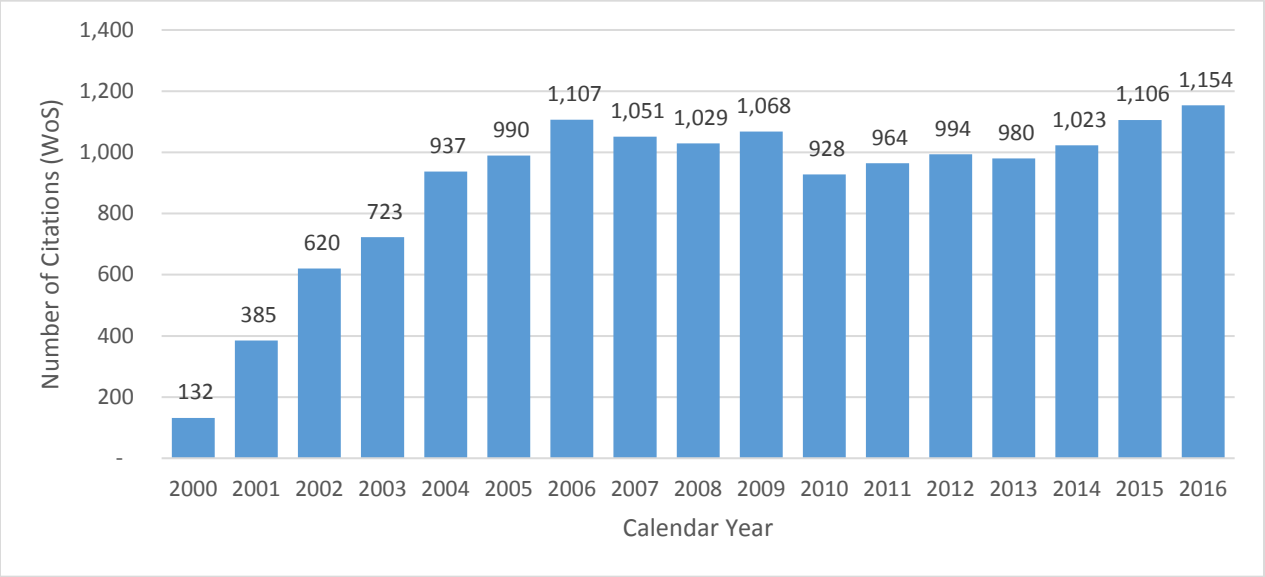

Figure 1: Citations to "The Protein Data Bank"

[2] Identification of commercial products does not imply recommendation or endorsement by the National Institute of Standards and Technology.

"The Protein Data Bank" has been cited by a wide range of articles, some of them very highly cited themselves. The citing article with the highest number of citations is "Coot: model-building tools for molecular graphics," with 14,709 citations. This article, along with five other highly cited articles, was published in *Acta Crystallographica Section D: Biological Crystallography*. Table 1 shows the top 10 papers citing the "The Protein Data Bank" article. Its citing by highly cited articles further increases the visibility and impact of "The Protein Data Bank."

Table 1. Top 10 Papers Citing "The Protein Data Bank" (By Number of Citations)

Citing Article	Number of Citations
Coot: modelbuilding tools for molecular graphics *Acta Crystallographica Section D: Biological Crystallography*	14,709
UCSF chimera - A visualization system for exploratory research and analysis *Acta Crystallographica Section D: Biological Crystallography*	9,563
PHENIX: a comprehensive Python-based system for macromolecular structure solution *Acta Crystallographica Section D: Biological Crystallography*	6,692
MolProbity: all-atom structure validation for macromolecular crystallography *Acta Crystallographica Section D: Biological Crystallography*	3,850
Inference of macromolecular assemblies from crystalline state *Journal of Molecular Biology*	3,418
MUSCLE: a multiple sequence alignment method with reduced time and space complexity *BMC Bioinformatics*	2,832
Protein structure prediction on the Web: a case study using the Phyre server *Nature Protocols*	2,714
PHENIX: building new software for automated crystallographic structure determination *Acta Crystallographica Section D: Biological Crystallography*	2,544
Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions *Acta Crystallographica Section D: Biological Crystallography*	2,186
MolProbity: all-atom contacts and structure validation for proteins and nucleic acids *Nucleic Acids Research*	1,982

"The Protein Data Bank" is cited by a variety of institution types, most predominantly by academia, foreign research institutions, and industry (Figure 2). It is cited most frequently by academic institutions with 2,514 institutions citing it a total of 25,987 times. The foreign research institutions category ranks second with 1,188 institutions, and third is industry (US and foreign) with 958 institutions. Hospitals also accounted for a significant number of institutions with 224.

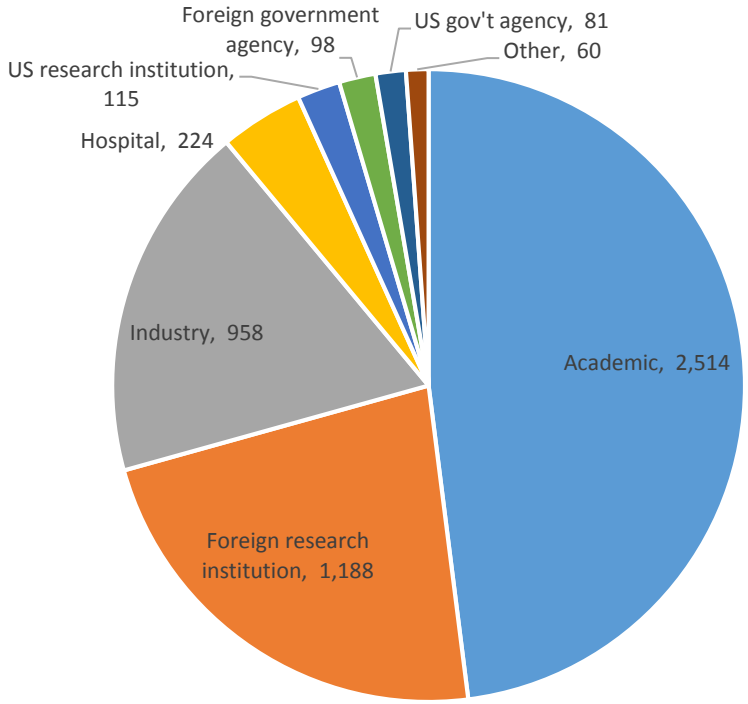

Figure 2: Institutions of Authors Citing "The Protein Data Bank"

"The Protein Data Bank" has far-reaching impact with papers from 151 different research areas citing the article. The area showing the highest impact, with 6,567 citations, is Biochemistry and Molecular Biology. Other research areas showing high impact include Biophysics (2,847 citations), Biochemical Research Methods (2,052 citations), and Computer Science Interdisciplinary Applications (1,830 citations) (Figure 3). The article has been cited by papers in areas ranging from Physical Chemistry to Dentistry, and Genetics/Heredity to Geology.

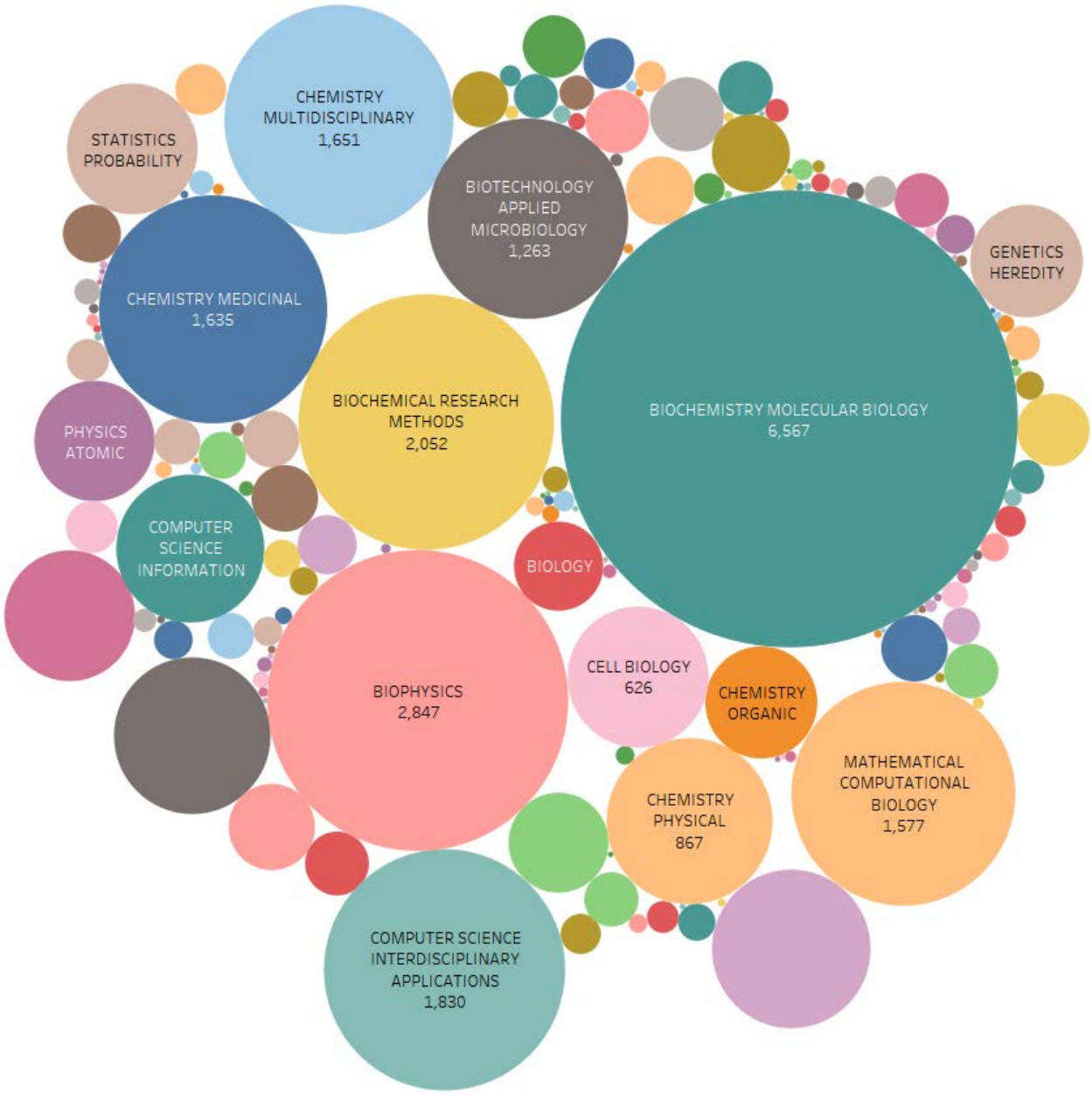

Figure 3. Research Areas Citing "The Protein Data Bank"

"The Protein Data Bank" has been cited by over 2,100 unique journal titles. Papers published in *Proteins: Structure, Function, and Bioinformatics*, with a 2015 Thomson Reuters impact factor of 2.499, cited "The Protein Data Bank" most often with 715 citations. Its impact factor places *Proteins: Structure, Function, and Bioinformatics* in the second quartile within its *WoS* subject category Biophysics. Papers published in *Nucleic Acids Research* have also cited "The Protein Data Bank" frequently with 571 citations. Its 2015 impact factor of 9.202 ranks it within the top 10% (18 out of 289) in its *WoS* subject category Biochemistry & Molecular Biology. Figure 4 shows the 25 journals citing "The Protein Data Bank" most frequently, giving the impact factor for each journal along with the number of citations to "The Protein Data Bank."

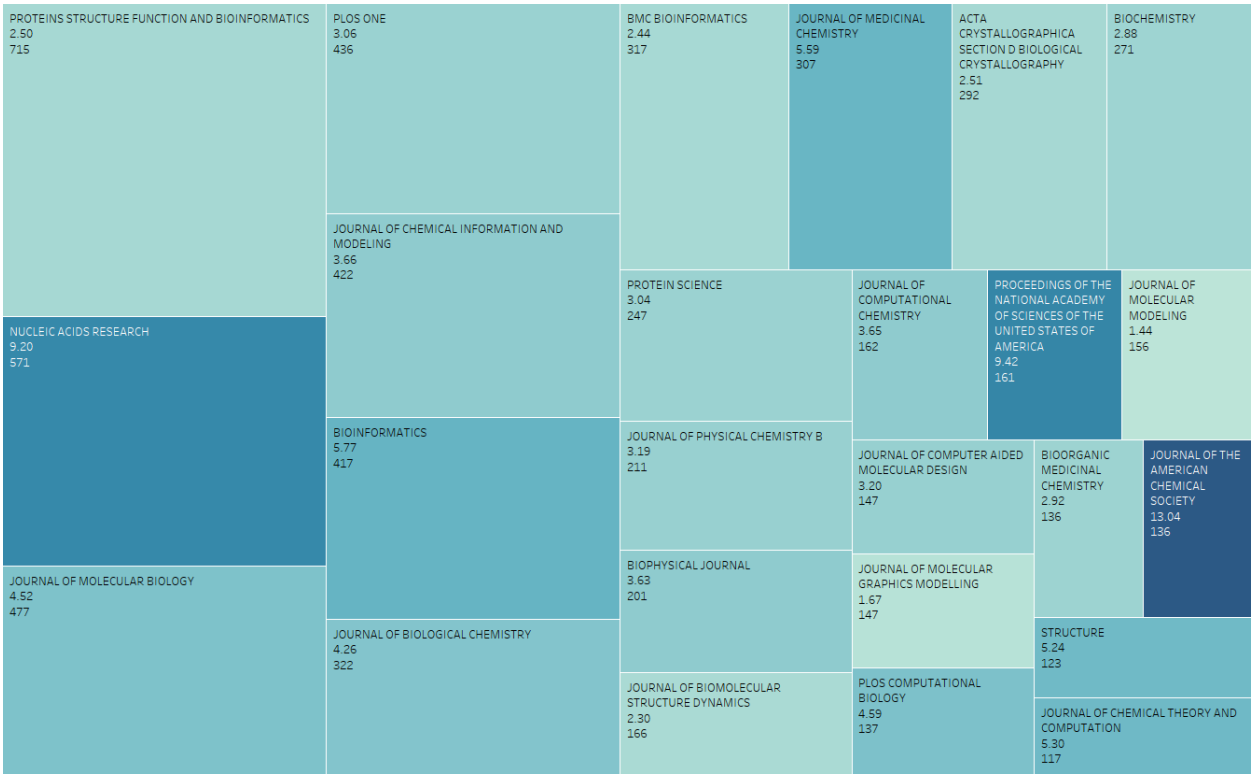

Figure 4. Top 25 Journals Citing "The Protein Data Bank" (Shaded by Impact Factor)

"The Protein Data Bank" has been cited by authors in 102 different countries and from every continent except Antarctica. It has been cited most frequently by authors in the United States (5,721 authors), Germany (1,437 authors), England (1,415 authors), and India (1,184 authors). Figure 5 shows the distribution of these citing authors around the world.

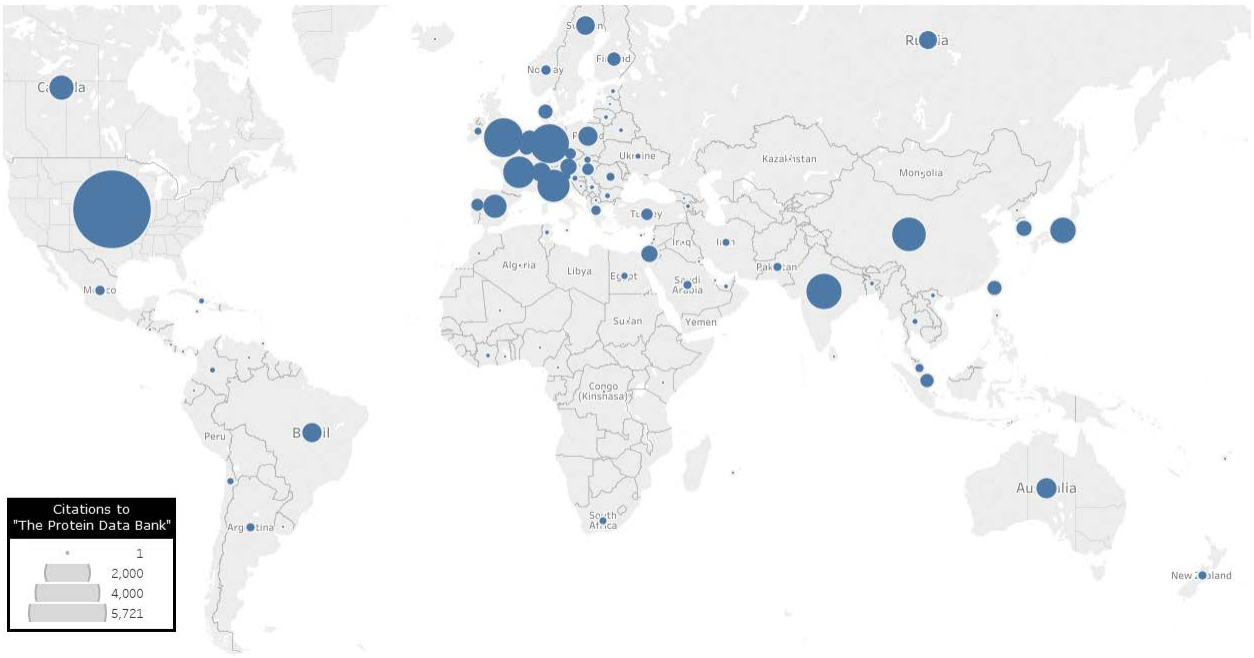

Figure 5. Countries of Authors Citing "The Protein Data Bank"

Conclusions and Recommendations

As the most highly cited NIST-authored article, "The Protein Data Bank," with over 15,000 citations, is considered a classic in its field. There are many ways to measure the impact of this article, as demonstrated through the various graphical representations shown in this paper. "The Protein Data Bank" has been cited across 151 different research areas in over 2,100 journals by authors from over 5,000 institutions in 102 countries. Too often, the impact of an article is measured simply by the number of citations, when in fact, a much richer story can be told through a closer look at the citation data.

Reference

Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T.N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. 2000. The Protein Data Bank. *Nucleic Acids Research* 28:1, 235–242, DOI: 10.1093/nar/28.1.235