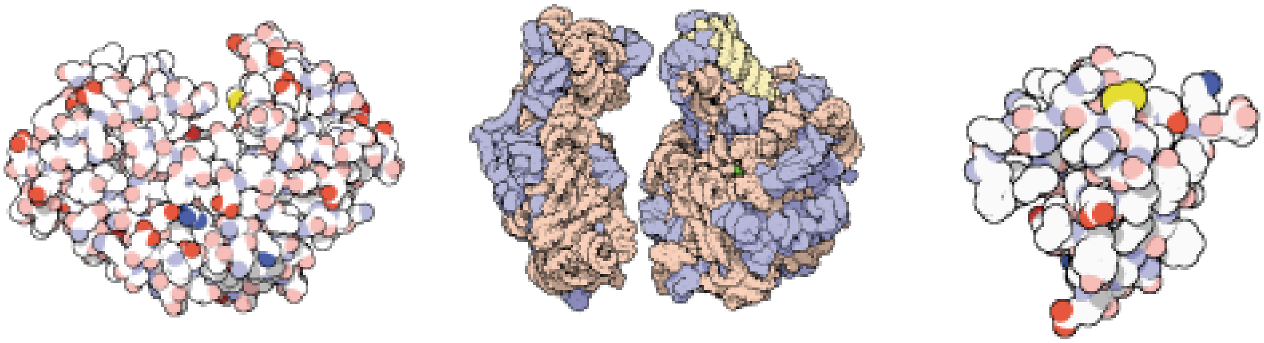
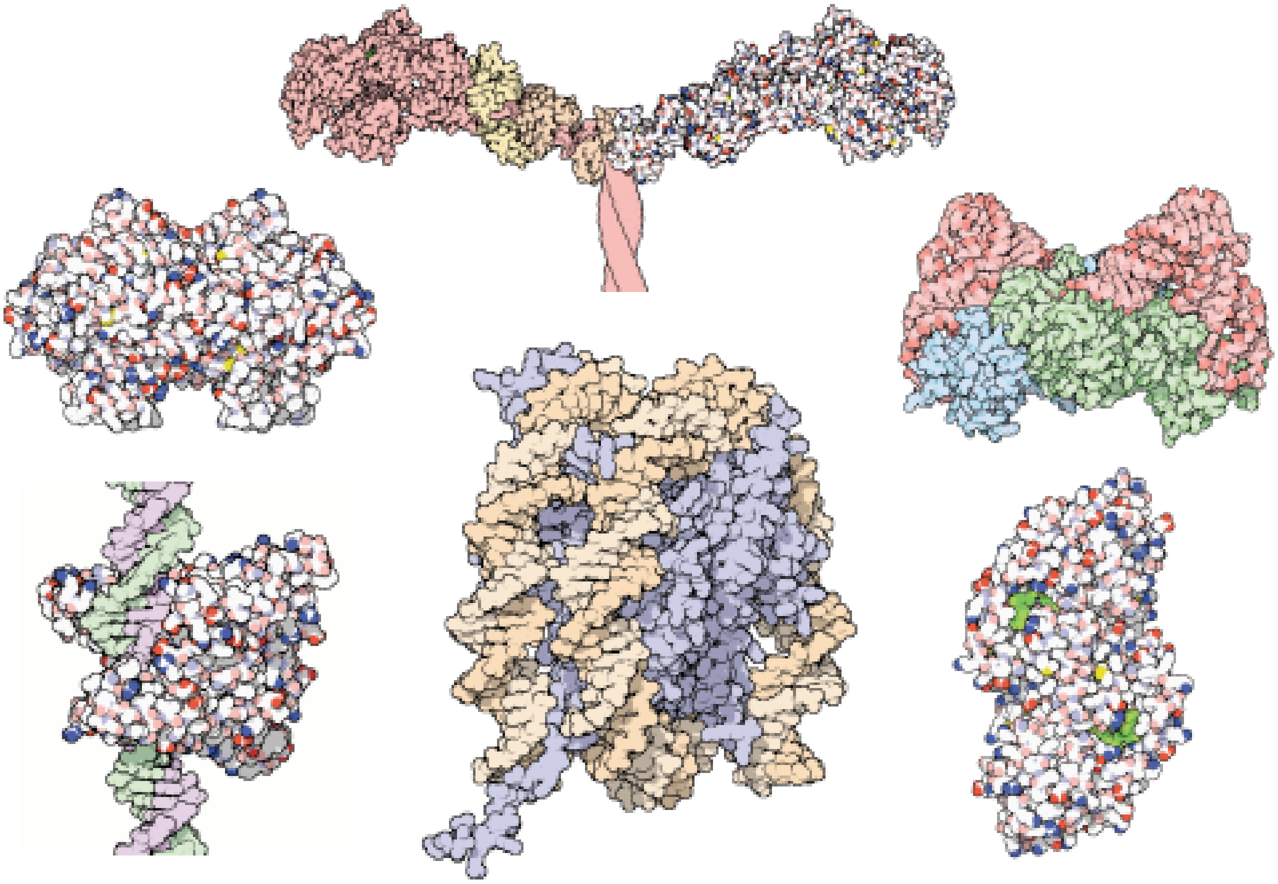


PROTEIN DATA BANK

PDB

ANNUAL REPORT

July 2000–June 2001



RESEARCH COLLABORATORY FOR STRUCTURAL BIOINFORMATICS

• Rutgers, The State University of New Jersey • National Institute of Standards and Technology •
• San Diego Supercomputer Center at the University of California, San Diego •

CONTENTS

RCSB PARTNER SITES

San Diego Supercomputer Center at the University of California, La Jolla, CA, USA

http://www.rcsb.org/pdb/
http://www.pdb.org/
ftp://ftp.rcsb.org/

Rutgers, The State University of New Jersey, Piscataway, NJ, USA

http://rutgers.rcsb.org/

National Institute of Standards and Technology, Gaithersburg, MD, USA

http://nist.rcsb.org/

OTHER RCSB MIRRORS

Cambridge Crystallographic Data Centre, United Kingdom

http://pdb.ccdc.cam.ac.uk/
ftp://pdb.ccdc.cam.ac.uk/rcsb/

National University of Singapore, Singapore

http://pdb.bic.nus.edu.sg/
ftp://pdb.bic.nus.edu.sg/pub/pdb/

Osaka University, Japan

http://pdb.protein.osaka-u.ac.jp/
ftp://ftp.protein.osaka-u.ac.jp/pub/pdb/

Universidade Federal de Minas Gerais, Brazil

http://www.pdb.ufmg.br/
ftp://vega.cenapad.ufmg.br/pub/pdb/

MESSAGE FROM THE DIRECTOR

It is a pleasure to introduce the second annual report of the Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank (PDB), which covers the period from July 1, 2000 through June 30, 2001. We have included in this document the many noteworthy accomplishments that have been made in the past year.

The PDB is the single international repository for three-dimensional structure data of biological macromolecules. It is an important resource for research in the academic, pharmaceutical, and biotechnology sectors, as well as a vital tool for education. Our mission is to provide the most accurate, well-annotated data in the most timely and efficient way possible to facilitate new discoveries and advances in science.

The PDB staff at the RCSB's three sites—Rutgers University, the National Institute of Standards and Technology (NIST), and the San Diego Supercomputer Center (SDSC)—has continued to work with skill and enthusiasm, and productive exchanges have occurred with colleagues and collaborators worldwide. Several highlights of our achievements during this time period include:

- Standardization of the PDB archive, and its release in mmCIF format.
- Efficient processing of increasingly complex entries, including ribosomal structures.
- New functionality and tools for query, such as a browser for enzyme classification and accurate source organism searches.
- Release of software tools that translate between mmCIF and PDB formats, create validation reports, and parse CIF formats.
- Support for CORBA standard for macromolecular structure.

In addition, our ongoing services, especially data deposition and annotation, data query, and data distribution, continue to be successful. The international PDB mirror sites have provided excellent access, and our active help desks allow us to be in constant contact with the community. The foundation set by these services allows us to further build upon this important resource. Our tools are in place to support new challenges. We are ready for the data from this era of structural genomics, and are constantly developing ways of getting these high throughput data in and out of the PDB efficiently.

As always, our future depends upon our user community. We look forward to your comments and suggestions so we can all continue to develop this resource.

Helen M. Berman
on behalf of the entire project team

Some members of the RCSB PDB Team (left to right): Phoebe Fagan, Dorothy Kegler, Haiyan Cheng, John Westbrook, Zukang Feng, Phil Bourne, Gary Gilliland, Diane Hancock, T.N. Bhat, Brad Kroeger, David Padilla, Victoria Colflesh, Helge Weissig, Narmada Thanki, Gnanesh Patel, Bohdan Schneider, Helen M. Berman, Nita Deshpande, Wolfgang Bluhm, Kyle Burkhardt, Lisa Iype, Ward Fleri, Christine Zardecki, Tammy Battistuz

WHAT IS THE PDB?

The three-dimensional structures of proteins and other biological macromolecules contained in the PDB are essential for avariety of research sectors. Understanding the shape of macromolecular structures aids in understanding how these molecules work. The PDB processes, stores, and disseminates structural coordinates and related information about proteins, nucleic acids, and protein-nucleic acid complexes. Some examples of the structures of interest that can be found in the PDB are DNA, RNA, viruses, and hemoglobin. The resource distributes information about all aspects of structural biology, including structural genomics, data representation formats, software and educational materials.

WHY IS IT IMPORTANT?

The three-dimensional structures of proteins and other biological macromolecules contained in the PDB are essential for a variety of research sectors as understanding the shape of macromolecular structures aids in understanding how these structures work. These structural data assist the pharmaceutical and biotechnology industries in understanding diseases and in identifying or developing drugs that can target diseases more accurately and with few or no side effects. Similarly, medical researchers gain new insight into causes, effects, and treatments that unlock the therapeutic potential of biological macromolecules, using the accurate, precise information in the PDB. To improve the quality of life on earth, scientists use PDB structural information in research directed at understanding the chemistry and biochemistry of natural processes. These efforts require the most consistent, well-annotated information available about the atomic structure of complex molecules.

New initiatives worldwide are focusing on structural genomics–structure determination in a high throughput mode to elucidate all structures in a given proteome, fill in protein fold space and enable a full understanding of a complete biochemical pathway. It is anticipated that these initiatives will produce a great influx of data. New ways to collect, validate, annotate, organize, view, and distribute data are necessary in order to meet the demands of managing and utilizing such a tremendous amount of information. The PDB will rise to this challenge through incorporation of the most recent technologies that facilitate the optimal methods to manage structural data.

HOW IS THE PDB MANAGED?

The PDB is managed by the Research Collaboratory for Structural Bioinformatics (RCSB), a nonprofit consortium of investigators and experts at three institutions: Rutgers, the State University of New Jersey; the National Institute of Standards and Technology (NIST); and the San Diego Supercomputer Center (SDSC), an organized research unit of the University of California, San Diego (UCSD). The RCSB is funded by the National Science Foundation (NSF), the Office of Biological and Environmental Research at the Department of Energy (DOE), and two units of the National Institutes of Health (NIH): the National Institute of General Medical Sciences (NIGMS) and the National Library of Medicine (NLM).

The RCSB project leaders manage the overall operation of the

PDB HOLDINGS AS OF JUNE 30, 2001

EXPERIMENTAL TECHNIQUE	MOLECULE TYPE				
	PROTEINS, PEPTIDES, AND VIRUSES	PROTEIN/NUCLEIC ACID COMPLEXES	NUCLEIC ACIDS	CARBOHYDRATES	TOTAL
X-RAY DIFFRACTION AND OTHER	11,630	556	580	14	12,780
NMR	1,918	73	383	4	2,378
THEORETICAL MODELING	288	19	23	0	330
TOTAL	13,836	648	986	18	15,488

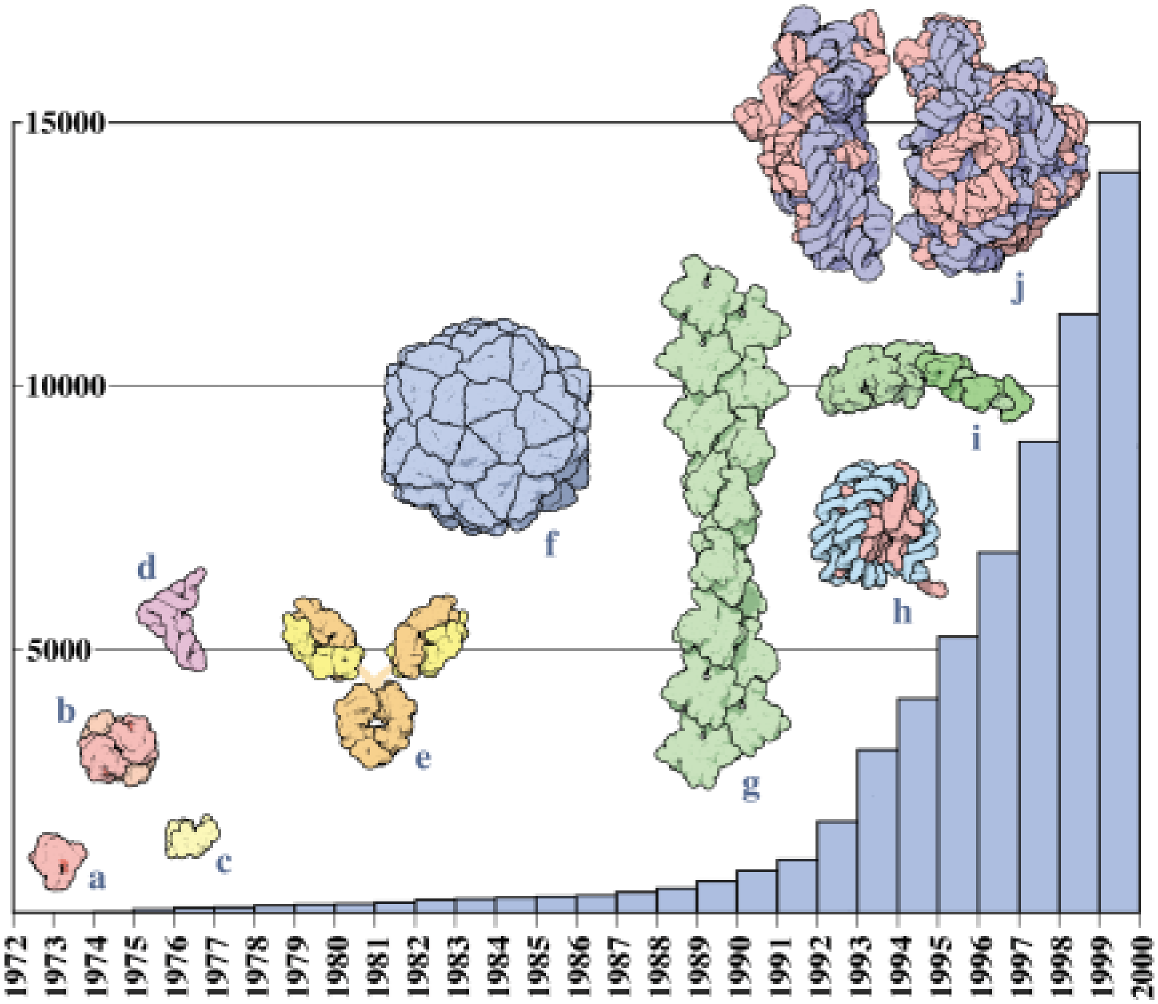
PDB. Dr. Helen M. Berman is the Director of the PDB and a Board of Governors Professor of Chemistry at Rutgers, the State University of New Jersey. Dr. Berman was part of the original team that developed the PDB at its inception at Brookhaven National Laboratory, and is the founder of the Nucleic Acid Database. Data deposition and processing are the responsibilities of the PDB team at RCSB-Rutgers, which is led by Dr. John Westbrook, Research Associate Professor of Chemistry. Data query and distribution functions are the responsibility of the PDB team at RCSB-SDSC, which is led by Dr. Philip E. Bourne, Professor of Pharmacology at UCSD and Senior Principal Scientist at SDSC. The exploration of issues relevant to NMR and management of the physical archive are the responsibilities of the PDB team at RCSB-NIST, which is led by Dr. Gary Gilliland, Chief of the Biotechnology Division of the Chemical Science and Technology Laboratory.

THE MISSION OF THE RCSB-PDB TEAM

The RCSB seeks to enable science worldwide by offering resources to improve the understanding of structure-function relationships in biological systems. The RCSB integrates production-level data and software resources, and it shares research results and software developments. It is our belief that new scientific advances will come from accurate, consistent, well-annotated three-dimensional structure data delivered in a timely and efficient way. In order to fulfill this mission, the capabilities of the PDB will continue to be significantly extended.

THE HISTORY OF THE PDB

The PDB was established containing 7 structures at Brookhaven National Laboratory in 1971.[21] Full responsibility for the operation and enhancement of the PDB was transferred to the RCSB on July 1, 1999,[22] at which time the PDB contained data for more than 10,000 structures.



Growth chart of the PDB that highlights example structures from different time periods: ***a.*** *myoglobin,*[1,2] ***b.*** *hemoglobin,*[3,4] ***c.*** *lysozyme,*[5,6] ***d.*** *transfer RNA,*[7-10] ***e.*** *antibodies,*[11,12] ***f.*** *entire viruses,*[13] ***g.*** *actin,*[14] ***h.*** *the nucleosome,*[15] ***i.*** *myosin,*[16] *and* ***j.*** *ribosomal subunits.*[17-19] *Images were created by Dr. David Goodsell, who creates the PDB's Molecule of the Month series. Figure originally appeared in the* International Union of Crystallography Newsletter.[20]

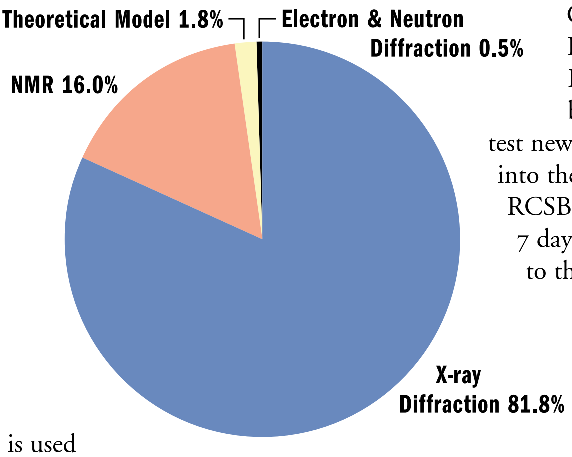
HOW DOES IT WORK?

The PDB is an important biological database. Currently in an average month, approximately 260 structures are deposited, 200 structures are released, and 2.6 million files of individual structure entries are downloaded from the PDB.

DATA INPUT

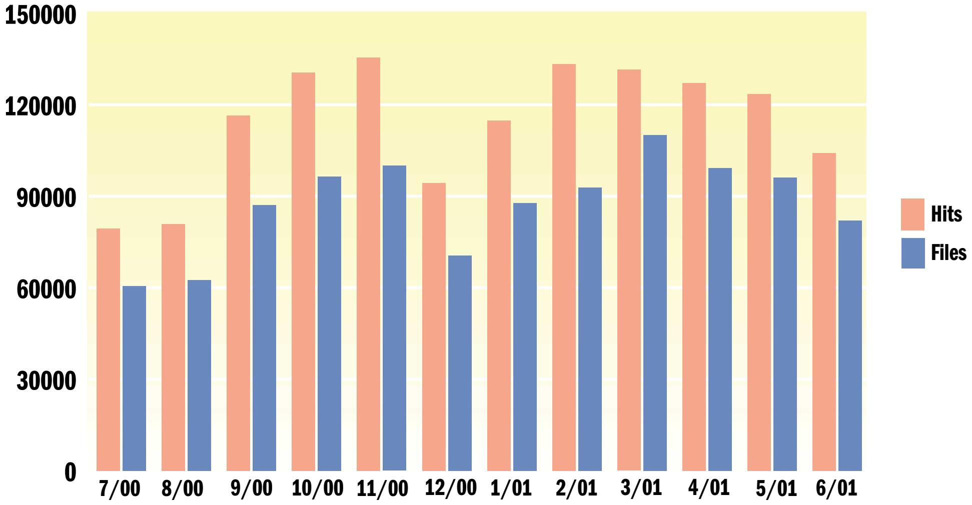
A key component of creating the public archive of information is data processing, which is the efficient capture and curation of data. The entire process consists of data deposition, validation, and annotation. Data from experiments using X-ray crystallography, nuclear magnetic resonance (NMR), and other methods are deposited to the PDB. Data are deposited using the AutoDep Input Tool (ADIT), which is available on-line from sites at RCSB-Rutgers (US) and the Institute for Protein Research (Japan). ADIT is also used to process data at these sites. Data are also accepted via FTP and e-mail, and then processed and annotated using ADIT. Structures may also be deposited to the PDB using the AutoDep system at the European Bioinformatics Institute (EBI); these data are processed at the EBI and forwarded to the RCSB for release.

Experimental Source for ADIT Depositions (July 2000-June 2001)



ADIT is built on top of a program that is used for data validation and exchange between different formats, and the macromolecular Crystallographic Information File (mmCIF) dictionary (**http://deposit.pdb.org/mmcif/**). mmCIF is an ontology of 1,700 terms defining macromolecular structure and related experiments.[23] After checks are performed by PDB staff, validation reports and a completed PDB file are returned to the depositor for review. Depositors also have the option to independently perform these checks using validation software released by the PDB. When finalized, the complete entry, including its status information and PDB ID, is loaded into the core relational database. This entire process is completed by the PDB staff with an average turnaround of less than two weeks.

Daily Average WWW Access by Month for the Primary RCSB Site (www.pdb.org)

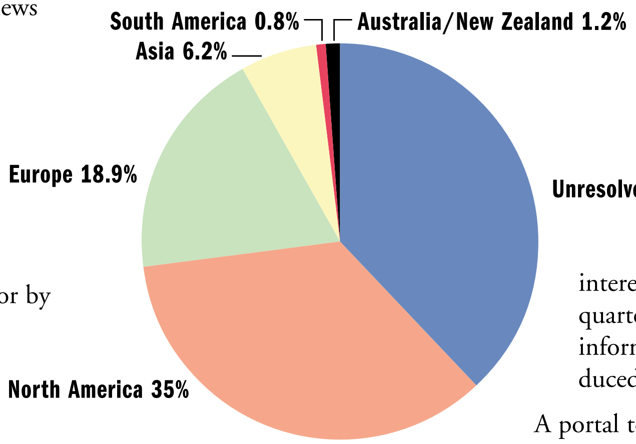


DATA DISTRIBUTION AND ACCESS

The PDB is a free service available through the Internet. The main PDB Web site at SDSC receives an average of more than 100,000 hits per day from all over the world—more than one hit per second, 24 hours per day, seven days per week. Additionally, there are six RCSB PDB mirror sites around the world at RCSB-Rutgers (US), RCSB-NIST (US), Osaka University (Japan), the National University of Singapore (Singapore), the Cambridge Crystallographic Data Centre (United Kingdom), and the Universidade Federal de Minas Gerias (Brazil). A beta Web site is available for users to test new features before they are incorporated into the main Web site and mirrors. All RCSB sites are maintained 24 hours per day, 7 days per week. New structures are added to the PDB holdings each Wednesday by 1:00 AM Pacific Time.

The PDB site offers several different interfaces to query the database. Entering the PDB ID of the target macromolecule performs the simplest search. These IDs are usually included in published papers describing the structure. A search by PDB ID produces the entry's Structure Explorer page. Each Structure Explorer page provides summary information about the entry, the atomic coordinates, derived geometric data, and experimental data (X-ray structure factors and NMR constraint data, where available). Structurally similar "neighbor" entries as computed using various methods are provided, along with options to further study aspects of the molecule, such as the secondary structure or primary amino acid sequence. Dynamic links to the structure's entry in other databases are provided by the Molecular Information Agent (MIA), and are accessible under the Other Sources option of the Structure Explorer page. Views of the structure are provided as static images, and in VRML, RasMol,[24] MICE,[25, 26] Chime,[27] and QuickPDB (Java). Multiple structures can be retrieved by using the keyword search functionality on the home page, by using the SearchLite interface, or by using the customizable SearchFields interface that searches on parameters selected by the user. The resulting Query Result Browser lists all molecules that meet the user's query specifications, and allows for exploration of one or more of the resulting structures. Options to refine the query or create tabular reports from such results are also available. A PDB or mmCIF format file for any structure can be downloaded as plain text or in one of several compression formats from the PDB Web site. Files may also be downloaded from the PDB FTP server.

PDB Hits to www.pdb.org Based on Region
38,034,462 total (July 2000–June 2001)

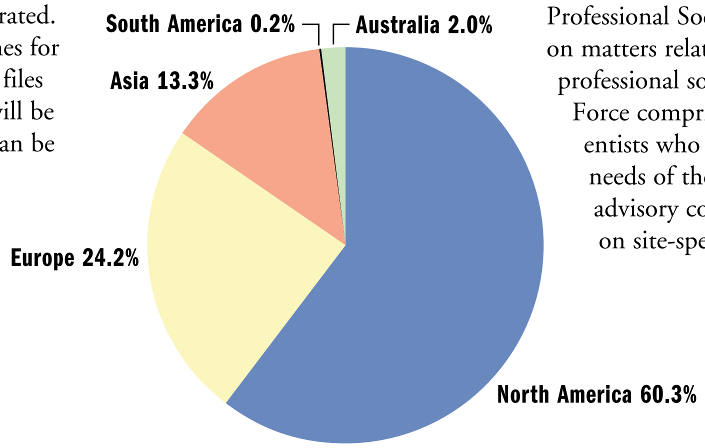


PHYSICAL ARCHIVE

The master physical archive, containing paper, magnetic, and electronic records, is maintained at the RCSB-NIST site. The installation of two automated filing systems is complete, and legacy paper files have been integrated. Documents can be retrieved by searches for author, PDB ID, or keyword. Legacy files currently stored on magnetic media will be read and stored on disk so that they can be more easily maintained and accessed.

A snapshot of the complete query and distribution production system is made by RCSB-SDSC and sent to RCSB-NIST for long term archiving each month.

ADIT Deposition Demographics
(July 2000–June 2001)



OUTREACH AND EDUCATION

The PDB promotes an active dialog with its user community to provide information about the resource, gain feedback, and provide materials for a broader understanding of structural biology. This is achieved in part through accessibility—the PDB maintains an active help desk and has a strong presence at meetings through presentations, user meetings, and exhibit booths. The **info@rcsb.org** electronic help desk generally responds to inquiries within a day or two. The RCSB-Rutgers site also maintains two other addresses for user support: **deposit@rcsb.rutgers.edu**, for general deposition and processing questions; and **help@rcsb.rutgers.edu**, for ADIT information. In addition, the **pdb-l@rcsb.org** listserv facilitates exchanges among members of the PDB community.

The PDB Web site is updated weekly with news, recent developments, and improvements. Newsletters detailing the latest enhancements to the PDB and other issues of interest to the user community are published quarterly in electronic and paper form. Other informative flyers and materials are also produced and circulated.

A portal to educational resources is available with links to various tutorials and videos for students and teachers of molecular biology. Highlighted on the PDB home page each month is Dr. David Goodsell's Molecule of the Month column, a feature intended for a general audience that focuses on a key biological molecule.

ADVISORY COMMITTEES

The PDB has continued to solicit the advice of several committees. The eleven members of the PDB Advisory Committee, chaired by Dr. Stephen K. Burley, are a team of experts in X-ray crystallography, NMR, modeling, and bioinformatics from institutions around the globe. A Database Advisory Committee includes six directors of other international data resources. The Professional Societies Committee advises on matters related to our interactions with professional societies. The NMR Task Force comprises eleven distinguished scientists who provide guidance on the needs of the NMR community. Local advisory committees are also consulted on site-specific matters.

PROGRESS AND ACHIEVEMENTS

A COLLECTION OF HIGHLIGHTS

DATA UNIFORMITY: RELEASE OF A STANDARDIZED ARCHIVE

One PDB objective is to make the archive as consistent and error-free as possible. Improvements in experimental methods, functional knowledge of proteins, and methods used to process these data have introduced various inconsistencies into the archive that limit the accuracy of queries. The PDB's Data Uniformity Project enhances the consistency of existing (legacy) entries and maintains a consistent method of annotating current depositions. During this period, the PDB archive has been standardized and released in mmCIF format.

As part of the Data Uniformity project, all files have been rechecked for validity, and errors have been corrected for data items such as sequence-coordinate consistency and atom nomenclature for macromolecules and ligands. Specific records have been reviewed and remediated for parameters such as the inclusion of synonyms and names used by other data centers. Molecular names and synonyms for all chains in the PDB are now available.

All legacy PDB entries and the recent RCSB entries are available in mmCIF format from the PDB beta FTP site at **ftp://beta.rcsb.org/pub/pdb/uniformity/data/mmCIF/**. The files follow the latest version of the mmCIF dictionary supplemented by an exchange dictionary developed by the PDB and the EBI. This exchange dictionary can be obtained from **http://deposit.pdb.org/mmcif/**.

An application program called CIFTr was made available for translating files in mmCIF format into files in PDB format. CIFTr works on UNIX platforms, and can be downloaded at **http://deposit.pdb.org/software/** (see below for more information).

ENHANCEMENTS TO DEPOSITION AND PROCESSING

From July 2000–June 2001, 3148 files were deposited with the PDB, and the average complexity of the molecules (number of amino acid residues) increased about 20 percent. Depositions containing over 100,000 atoms were processed and released. Some statisticxs on these data are shown in the charts on pages 4-5.

The ADIT interfaces for X-ray and NMR deposition now include more detailed information, dictionary enhancements, and expanded help systems. The values have been updated for several pull-down menus. In response to comments from users, depositors may now prerelease sequence data ahead of structure coordinate data.

The ADIT deposition and annotation site established at the Institute for Protein Research at Osaka University in Osaka, Japan has been operational for a year. Entries deposited at this site have been processed by staff at the Laboratory of Protein Informatics and are incorporated into the PDB archive. Part of the success of this cooperative agreement is due to the productive visits that the RCSB and Osaka group members have made to both sites. We look forward to continuing collaborations.

We have also continued collaborating with EBI in data deposition and processing, as well as determining requirements for the efficient capture of cryogenic electron microscopy (cryo-EM) data.

ENHANCEMENTS TO QUERY AND REPORTING

A new interface can show how many structures exist at some level in the Enzyme Commission (EC) hierarchy. This required the accurate assignment of enzyme numbers to all relevant PDB structures and integration of EC nomenclature into the database system.

Sequences can be accessed in advance of structure coordinate release. As available through the PDB's status search, this new capability permits valid tests of structural prediction algorithms. Users may query all available sequences, or query based on criteria such as title or deposition date.

A recent visit with Kyle Burkhardt (RCSB-Rutgers) and the Osaka group at the Institute for Protein Research, Osaka University, Japan.

Back row: Reiko Igarashi, Takashi Kosada, Kyle Burkhardt, Yumiko Kengaku

Front row: Genji Kurisu, Masami Kusunoki

Several new searching functions have been released. It is now possible to search by the number of chains on a structural backbone or in a complex. Users can also search by source organism, including synonyms and common names, so that searches on "human" and "*Homo sapiens*" return the same entries. The keyword search function now performs both exact and partial word matching, and it is possible to query on the titles of entries.

Further developments include the integration of the Molecular Interactive Collaborative Environment (MICE) viewer with custom visualization options. Researchers located around the world can now view the same structure simultaneously and pass control of rendering and motion from participant to participant.

Users can now customize tabular reports or use the preformatted tables available, and there are new custom display options in the Search Fields interface. Furthermore, the Sequence Details section of the Structure Explorer page now points to entries in the major sequence databases for the structure explored.

ENHANCEMENTS TO THE PDB WEB DESIGN

The Web site was reviewed for usability by the PDB staff and community. The home page has been revised to emphasize mirror sites, to permit both keyword and PDB ID searching, and to improve access to documents, format descriptions, and other materials. All of these enhancements were included in a revised tutorial on the query functions of PDB, linked through the home page. Upgrades were also made to the help documentation for the View Structure section of the Structure Explorer, with links to download plug-ins for VRML, RasMol,[24] Chime,[27] and MICE.[25,26]

Remediated files, related software, and update notices are archived at the Data Uniformity Project Web page at **http://www.rcsb.org/pdb/uniformity/index.html**.

The PDB has compiled a variety of structural genomics links at **http://www.rcsb.org/pdb/strucgen.html**. The purpose of this page is to provide an entry point to additional information on structural genomics relevant to PDB users.

GROWTH IN ACCESS AND DISTRIBUTION

Usage of the primary PDB Web site, its mirrors, the FTP site, and beta test site continues to grow. Monthly averages of Web site hits at the primary site have totaled more than 100,000 per day for the first six months of 2001, and approximately 87,000 downloads of files were made each month. Access to the primary PDB Web site and FTP site were enhanced by the addition of an alternative Internet Service Provider, increasing bandwidth to 40 Mbit/s. Redundant, load-balanced systems now serve these sites, supporting faster and more reliable service, greater throughput, and the ability to handle more users. It is clear that the PDB is a global resource whose growth requires intelligent anticipation of future directions in structural biology and related fields.

NEW SOFTWARE RELEASED—CIFTr, VALIDATION SOFTWARE, AND A CIF PARSER

CIFTr, a tool used by PDB staff in data processing that translates files from mmCIF to PDB format, has been released publicly. The program works on UNIX platforms, and can be downloaded at **http://deposit.pdb.org/software/**. It also provides the option of producing a file with a blank chain ID field for structures with a single chain, and the option of producing files with standard IUPAC hydrogen nomenclature for standard L-amino acids.

The validation software used by ADIT to run checks on structures as a part of primary data processing and as part of data uniformity has been compiled into a suite of programs available for download. Designed to work with files in mmCIF or PDB format, the beta version of this validation software can be downloaded in binary form for SGI, SUN, and Linux platforms from **http://deposit.pdb.org/software/**. Reports produced include an Atlas entry, a summary report, and a collection of structural diagnostics including bond distance and angle comparisons, torsion angle comparisons, base morphology comparisons (for nucleic acids), and molecular graphic images. In addition, reports from PROCHECK,[28] NUCheck,[29] and SFCHECK[30] are also made available.

A set of simple object-oriented Perl modules and scripts for parsing STAR (Self-defining Text Archive and Retrieval format)-compliant files and dictionaries, such as mmCIF, were released. Users with a working knowledge of Perl and a basic familiarity of CIF or other STAR-compliant data file formats will benefit from these tools. The released scripts are a mixture of basic utility scripts and very simplistic examples that are meant to test certain methods in the modules. Users can also write their own customized scripts. To download these modules or for more information, please refer to the documentation at **http://pdb.sdsc.edu/STAR/**.

CORBA

Of particular importance to the RCSB's efforts to provide greater access was the decision of the Board of Directors of the Object Management Group (OMG), on February 27, 2001, to adopt the Common Object Request Broker Architecture (CORBA) Macromolecular Structure Specification. Closely aligned with the mmCIF standard, this specification opens the door to more seamless and specific access to PDB data by providing a standard application programming interface (API) that will allow direct access by remote programs to the binary data structures of the PDB. Investigators worldwide will be able to retrieve a single data item from an entire PDB file for use in a local application, without having to download the entire file. The RCSB has become a member of OMG, which oversees the development of many other open standards for object-oriented computing in the life sciences. Collectively, these specifications will provide a robust framework for integration of key data resources needed by the structural biology community.

DEVELOPMENTS IN NMR

The PDB is working to improve structure deposition and annotation services for data acquired from NMR experiments, and continues to work with our NMR Task Force, the Collaborative Computational Project for NMR (CCPN), and the BioMagResBank (BMRB). Meetings with BMRB have resulted in the development of a plan to complete the NMR data dictionary, and the creation of a prototype integrated deposition tool.

CD-ROM

The PDB distributes a quarterly CD-ROM snapshot of its holdings. Four sets were created and released during the period of this report, at no cost to the user. The CD-ROM subscription list has doubled in the past year. New technologies for distributing this data are being explored as the growth of the archive increases.

OUTREACH AND EDUCATION

The PDB frequently interacts with the user community through a variety of mediums. PDB sponsored exhibit booths at the American Crystallographic Association's (ACA) Annual Meeting (July 2000), the 14th Symposium of the Protein Society (August 2000), the Eighth International Conference on Intelligent Systems for Molecular Biology (ISMB, also in August 2000), the Biophysical Society annual meeting (February 2001), and PITTCON (March 2001). PDB members presented talks and demonstrations at more than 30 meetings around the world. User group meetings were held at the ACA meeting and locally at RCSB sites.

The PDB helped to develop a tutorial on the role of hemoglobin as part of the Envision, Explore, Engage project of the National Partnership for Advanced Computational Infrastructure. The tutorial is part of the Molecular Science CD-ROM available from **http://e3.sdsc.edu/**.

Additions to the Web site included a compendium of mmCIF resources, a portal to information about structural genomics, and updates regarding the data uniformity project. Other PDB Web portals, such as the educational resources section, are updated frequently with new features.

The journal Science Watch, published by the Institute for Scientific Information, noted that the main reference for the PDB[22] ranked second among the "Red Hot Research Papers of 2000" in terms of citations received,[31] and an article by Jeremy Cherfas in the succeeding issue praised the progress made by the PDB.[32] Other news about the PDB appeared in many printed periodicals and Web publications, including Science,[33] The New York Times,[34] Genome Biology,[35] Bioinform,[36] FEED[37] and the ACA and IUCr Newsletters. Finally, four issues of the PDB newsletter and the first RCSB PDB Annual Report for July 1999–June 2000 were produced and distributed.

PUBLICATIONS

H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N.Shindyalov, P.E. Bourne, (2000): The Protein Data Bank. *Nucleic Acids Research* **28** (1), pp. 235-242.

H.M. Berman, T.N. Bhat, P.E. Bourne, Z. Feng, G. Gilliland, H. Weissig, J. Westbrook (2000): The Protein Data Bank and the Challenge of Structural Genomics. *Nature Structural Biology* **7** (11), pp. 957-959.

T.N. Bhat, P.E. Bourne, Z. Feng, G. Gilliland, S. Jain, V. Ravichandran, B. Schneider, K. Schneider, N. Thanki, H. Weissig, J. Westbrook, H.M. Berman (2001): The PDB data uniformity project. *Nucleic Acids Research* **29** (1), pp. 214-218.

COLLABORATIONS WITH OTHER ORGANIZATIONS

New and ongoing collaborations include:

Our collaboration on NMR data deposition continues with the BioMagResBank (BMRB), stressing the development of a data dictionary and an integrated deposition system based on ADIT.

We continue to work with the Cambridge Crystallographic Data Centre (CCDC) on methods for ligand validation, and we are now mirroring ReLiBase+, a CCDC ligand resource.

Collaborations have begun with the Brazilian Agricultural Research Corporation (Embrapa) to make features of STING Millennium, a suite of Web-based programs for simultaneous analysis and display of structure and sequence, available to PDB users.

We worked with Emerald Biosystems on the special problems of handling structures with licensing restrictions, and that work continues.

We have worked with the European Bioinformatics Institute (EBI) on processing of depositions at the EBI site, and on a data exchange dictionary, which is undergoing testing.

Close collaborations have been maintained with the Institute for Protein Research at Osaka University, where ADIT is used for the deposition and processing of structures.

Our colleagues at the National Center for Biotechnology Information (NCBI) at NIH are working with us on ways to ensure that PDB files can be used by the NCBI-developed databases.

We are working with Dr. Alexander Wlodawer and Dr. Jiri Vondresek to move the HIV Protease Database from NCI in Frederick, Maryland, to NIST, incorporating uniformity-compliant PDB file data. The database has been converted to Oracle and the new version will soon be released for alpha testing.

Dr. David Goodsell of The Scripps Research Institute continues to contribute the Molecule of the Month feature to the PDB site, and his intricate drawings of molecules continue to delight and educate PDB users.

Other collaborators include Dr. Paul Adams (Lawrence Berkeley National Laboratory), Dr. Wladek Minor (University of Virginia) and Dr. Zbyszek Otwinowski (University of Texas) on developing software for structural genomics, Dr. Alexei Adzhugei (Swiss Bioinformatics Institute and GlaxoSmithKline) on a models and mmCIF database project, as well as Anne Kuller (BioSync), Dr. Peter Karp (Metacyc), Dr. Ernest Laue (CCPN), Dr. Dietmar Schomburg (BRENDA), and Dr. Cherri Pancake (Oregon State University).

THE FUTURE OF THE PDB

The PDB has made significant progress in achieving many goals and will continue to improve our services to the community by adding new functionality to the PDB and by assessing and anticipating the needs of a growing user base.

DATA DEPOSITION AND PROCESSING

Data deposition and processing procedures will continue to be streamlined and automated as much as possible. With the expected influx of data in the coming years, this effort will remain a high priority.

It is expected that improvements in technology and experimental methods will accelerate the growth of the number of structures deposited into the PDB holdings. By some estimates, the archive could grow to approximately 35,000 structures by 2005, which would nearly triple its size in less than five years. Structural genomics will be the major catalyst of this growth. With new methods for data gathering, the need to create new data items is apparent. Additionally, new ways to capture and facilitate access to this information need to be implemented. The PDB will also accommodate new requirements, such as new protocols for depositions of cryo-EM data and powder crystallography data. The PDB will also continue its collaborations with the NIH structural genomics projects and participation in structural genomics efforts worldwide.

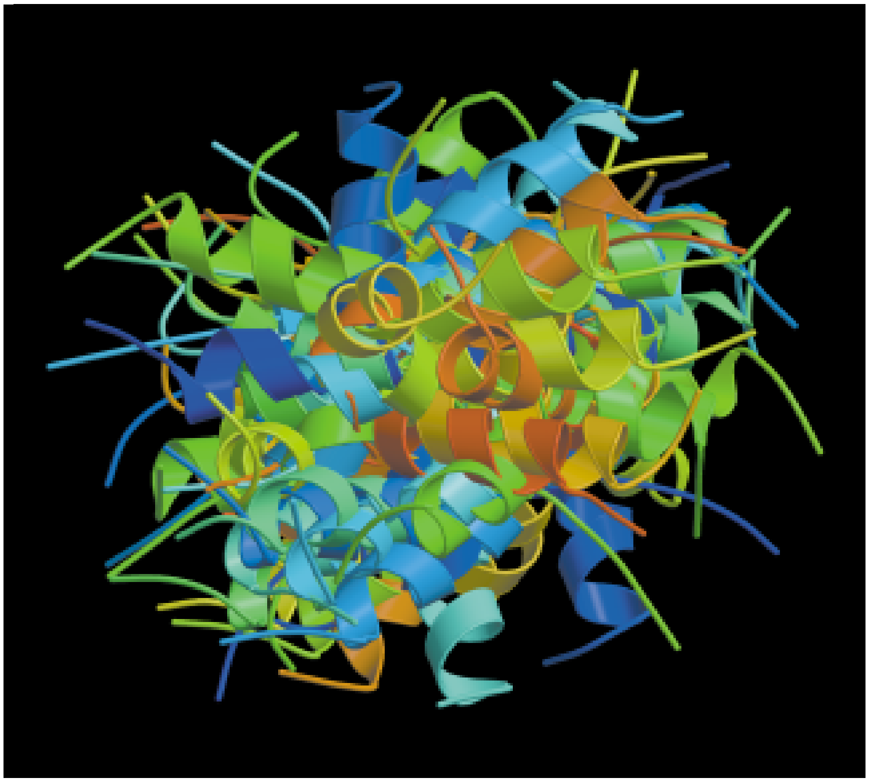
Release of data processing software tools is in progress, and will continue with a full release, including ADIT, in the future. Data deposition and distribution activities will be managed and integrated among partner sites. New tools for data processing that incorporate the mmCIF standard will be developed.

DEVELOPMENTS IN DATA ACCESS AND QUERY

The PDB will continue to develop the PDB Web site and its query capabilities. Additional data mining tools will be implemented, data exchange and database interoperability will be enhanced, and new resources such as STING Millennium and redundancy reduction capabilities, will be made available.

We anticipate the establishment of a seventh mirror site at the Max-Delbrück-Center for Molecular Medicine in Berlin in the near future.

The PDB will also continue its efforts to develop a standard application interface for macromolecular data based on CORBA and the mmCIF standard. Unlike current access in which users are required to retrieve and parse complete PDB files, an implementation of this CORBA API will allow applications to retrieve a single data item from a remote PDB server and import it for use in a local application.



The p53 tumor suppressor is a molecular watchdog that continually watches for damaged cells. Sensing DNA damage, it can halt cell division, ensuring that the cell does not pass on its faulty genetic material, or even initiate programmed cell death to remove the problem permanently. The structure in PDB entry **1hs5** *contains an altered form of the central domain of p53. This domain normally ties four separate p53 chains together, forming a tight tetramer that is needed for proper biological action. The form in this file has been mutated so that it only forms a dimer, which compromises some, but not all, of its functions. Mutations in p53 that block or modify its function often lead to the development of cancer.*

PDB ID: 1hs5

T.S. Davison, X. Nie, W. Ma, Y. Lin, C. Kay, S. Benchimol, C.H. Arrowsmith (2001): Structure and Functionality of a Designed p53 Dimer. J. Mol. Biol. *307, p. 605.*

CONTINUED DATA UNIFORMITY

The PDB will integrate the archive of PDB standardized data that has been released in mmCIF format with the relational database. We will continue to review the PDB archive in response to user input and by examining additional data items.

STRUCTURAL GENOMICS

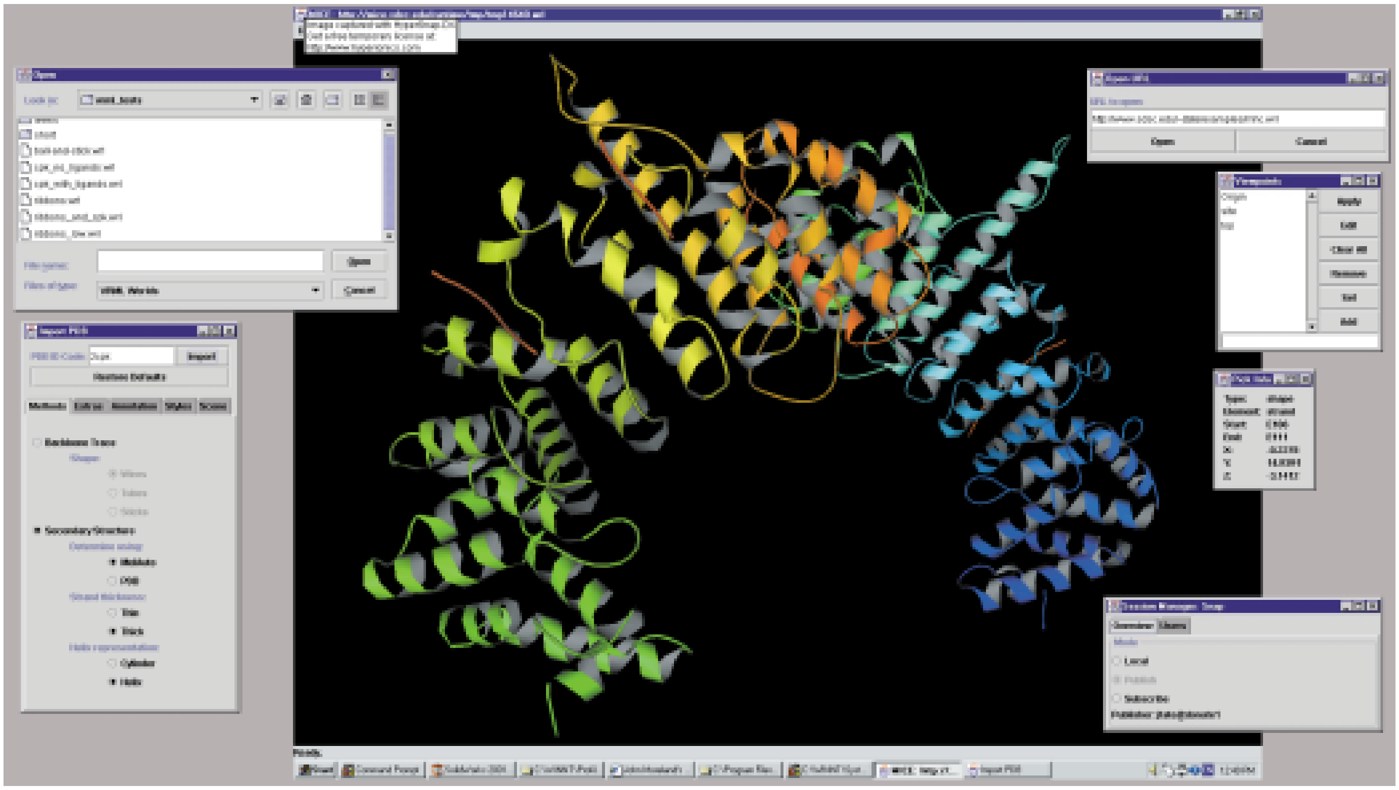
The PDB has been an active participant in the structural genomics initiatives. For these entries, PDB staff will collect information about the materials and methods used in structure determinations. The corresponding data items that describe such structures are now being developed in collaboration with structural genomics centers from around the world. Discussions with organizations such as BioSync, individual structural genomics projects, and synchrotron sources will facilitate the deposition of more detailed experimental data. Raw data collected at the Stanford Linear Accelerator Center (SLAC) are being archived at SDSC. PDB members will continue to be actively involved in structural genomics meetings and organizations, including membership on the International Task Force on Data Deposition, in order to discuss policies related to the deposition and release of these data.

FUTURE OF NMR

The PDB will continue its collaborations with the BMRB, the NMR Taskforce, and the CCPN software initiative. It is anticipated that there will be a prototype of the joint PDB-BMRB deposition system in the coming year.

OUTREACH AND EDUCATION

The PDB will remain engaged with its diverse user community of students, teachers, and researchers at future meetings and through its help services. Our newsletters and Web site will continue to develop. The CD-ROM distribution will continue on a quarterly schedule, and new ways of reducing the number of CDs in the set and their production costs will be explored.



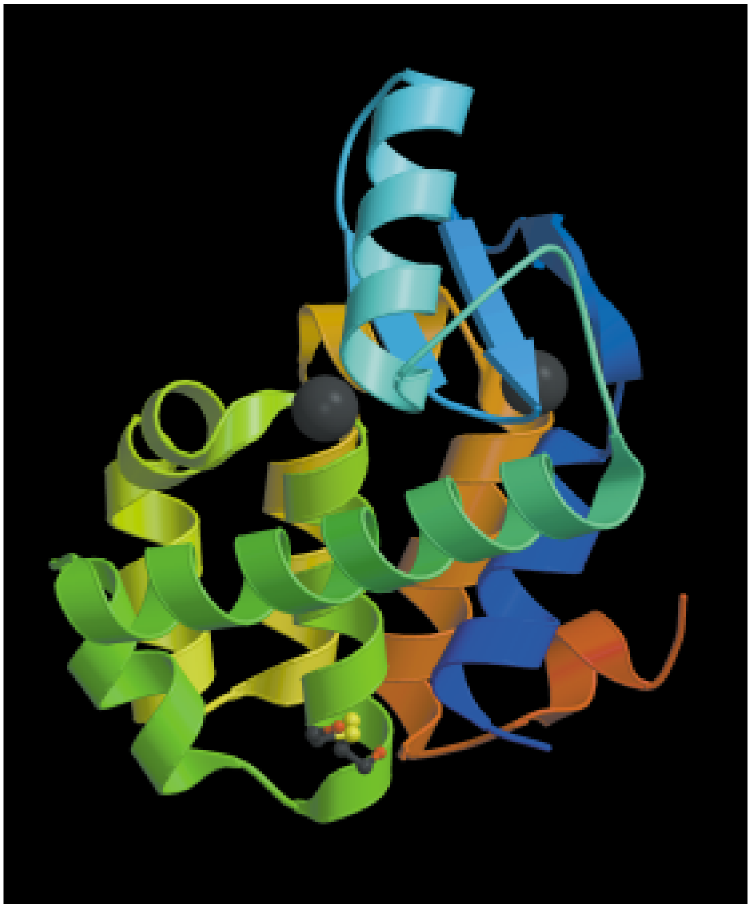
The MICE viewer provides a way for remote collaborators to examine and manipulate molecular images in real time. The screen shot above shows an example of a simple three-dimensional MICE scene the importin-beta Fxfg nucleoporin complex, represented in VRML.

PDB ID: 1f59

R. Bayliss, T. Littlewood, M. Stewart (2000): Structural Basis for the Interaction between Fxfg Nucleoporin Repeats and Importin-Beta in Nuclear Trafficking. Cell (Cambridge,Mass.) *102, p. 99.*

SELECTED REFERENCES

1. H.C. Watson (1969): The stereochemistry of the protein myoglobin. *Prog. Stereochem.* 4, p. 299.
2. J.C. Kendrew, G. Bodo, H.M. Dintzis, R.G. Parrish, H.Wyckoff (1958): A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature* **181**, pp. 662-666.
3. W. Bolton, M.F. Perutz (1970): Three dimensional fourier synthesis of horse deoxyhaemoglobin at 2.8 Å unites resolution. *Nature* **228**, pp. 551-552.
4. M.F. Perutz, M.G. Rossmann, A.F. Cullis, G. Muirhead, G. Will (1960): Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5 Å resolution. *Nature* **185**, pp. 416-422.
5. C.C.F. Blake, D.F. Koenig, G.A. Mair, A.C.T. North, D.C Phillips, V.R. Sarma (1965): Structure of hen egg-white lysozyme. A three dimensional Fourier synthesis at 2 Å resolution. *Nature* **206**, pp. 757-761.
6. C.C.F. Blake, L.N. Johnson, G.A. Mair, A.C.T. North, D.C. Phillips, V.R. Sarma (1967): Crystallographic studies of the activity of hen egg-white lysozyme. *Proc. R. Soc. London Ser. B* **167**, pp. 378-388.
7. B.E. Hingerty, R.S. Brown, A. Jack (1978): Further refinement of the structure of yeast T-RNA-Phe. *J. Mol. Biol.* **124**, pp. 523-524.
8. J.L. Sussman, S.R. Holbrook, R.W. Warrant, G.M. Church, S.-H. Kim (1978): Crystal structure of yeast phenylalanine transfer RNA. I. Crystallographic refinement. *J. Mol. Biol.* **123**, pp. 607-630.
9. F. Suddath, G. Quigley, A. McPherson, D. Sneden, J. Kim, S. Kim, A. Rich (1974): Three-dimensional structure of yeast phenylalanine transfer RNA at 3.0 Ångstroms resolution. *Nature* **248**, pp. 20-24.
10. J.D. Robertus, J.E. Ladner, J.T. Finch, D. Rhodes, R.S. Brown, B.F.C. Clark, A. Klug (1974): Structure of yeast phenylalanine tRNA at 3 Å resolution. *Nature* **250**, pp. 546-551.
11. J. Deisenhofer (1981): Crystallographic refinement and atomic models of a human Fc fragment and its complex with fragment B of protein A from *Staphylococcus aureus* at 2.9 and 2.8 Å resolution. *Biochemistry* **20**, pp. 2361-2370.
12. Y. Satow, G.H. Cohen, E.A. Padlan, D.R. Davies (1986): Phosphocholine binding immunoglobulin Fab McPC603. An X-ray diffraction study at 2.7 Å. *J. Mol. Biol.* **190**, pp. 593-604.
13. T.A. Jones, L. Liljas (1984): Structure of satellite tobacco necrosis virus after crystallographic refinement at 2.5 Å resolution. *J. Mol. Biol.* **177**, pp. 735-767.
14. W. Kabsch, H.G. Mannherz, D. Suck, E.F. Pai, K.C. Holmes (1990): Atomic structure of the actin:DNase I complex. *Nature* **347**, pp. 37-44.
15. K. Luger, A.W. Mader, R.K. Richmond, D.F. Sargent, T.J. Richmond (1997): Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **389**, pp. 251-260.
16. A. Houdusse, A.G. Szent-Gyorgyi, C. Cohen (2000): Three conformational statues of scallop S1. *Proc. Nat. Acad. Sci. USA* **97**, pp. 11238-11243.



Lysozyme, the first enzyme to ever have its structure solved (see ref. 5,6), is the most common protein in the PDB. It is a small, stable enzyme, making it ideal for research into protein structure and function. Lysozyme attacks the protective cell walls of bacteria, protecting us from the ever-present danger of infection.

PDB ID: 1g06

J. Xu, W.A. Baase, M.L. Quillin, E.P. Baldwin, B.W. Matthews (2001): Structural and Thermodynamic Analysis of the Binding of Solvent at Internal Sites in T4 Lysozyme. Protein Sci. ***10**, p. 1067.*

17. N. Ban, P. Nissen, J. Hansen, P.B. Moore, T.A. Steitz (2000): The complete atomic structure of the large ribosomal subunit at a 2.4 Å resolution. *Science* **289**, pp. 905-920.

18. F. Schluenzen, A. Tocilj, R. Zarivach, J. Harms, M. Gluehmann, D. Janell, A. Bashan, H. Bartels, I. Agmon, F. Franceschi, A.Yonath (2000): Structure of functionally activated small ribosomal subunit at 3.3 Å resolution. *Cell* **102**, pp. 615-623.

19. B.T. Wimberly, D.E. Brodersen, W.M. Clemons Jr., R. Morgan-Warren, A.P. Carter, C. Vonrhein, T. Hartsch, V. Ramakrishnan (2000): Structure of the 30S ribosomal subunit. *Nature* **407**, pp. 327-339.

20. (2001): International Union of Crystallography Newsletter **9** (1).

21. F.C. Bernstein, T.F. Koetzle, G.J. Williams, E.E. Meyer, M.D. Brice, J.R. Rodgers, O. Kennard, T. Shimanouchi, M. Tasumi (1977): Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, pp. 535-542.

22. H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne (2000): The Protein Data Bank. *Nucleic Acids Res.* **28**, pp. 235-242.

23. P.E. Bourne, H.M. Berman, K. Watenpaugh, J. Westbrook, P.M.D. Fitzgerald (1997): The macromolecular Crystallographic Information File (mmCIF). *Meth. Enzymol.* **277**, pp. 571-590.

24. R. Sayle, E.J. Milner-White (1995): RasMol: biomolecular graphics for all. *Trends Biochem. Sci.* **20**, p. 374.

25. P.E. Bourne, M. Gribskov, G. Johnson, J. Moreland, H. Weissig (1998): *Pacific Symposium on Biocomputing.*

26. J.G. Tate, J. Moreland, P.E. Bourne (1999): MSG (Molecular Scene Generator): a Web-based application for the visualization of macromolecular structures. *Journal of Applied Crystallography* **32**, pp. 1027-1028.

27. Chime. MDL Information Systems, Inc. United States.

28. R.A. Laskowski, M.W. McArthur, D.S. Moss, J.M. Thornton (1993): PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.* **26**, pp. 283-291.

29. Z. Feng, J. Westbrook, H.M. Berman (1998): NUCheck. Technical Report NDB-407. (Rutgers University, New Brunswick, NJ).

30. A.A. Vaguine, J. Richelle, S.J. Wodak (1999): SFCHECK: a unified set of procedures for evaluating the quality of macromolecular structure-factor data and their agreement with the atomic model. *Acta Crystallogr.* **D55**, pp. 191-205.

31. (2001): The hottest research of 1999-2000. *Science Watch®* **12** (2), pp. 1-2.

32. J. Cherfas (2001): Whether cited or not, protein databases proving popular. *Science Watch®* **12** (3), p. 8.

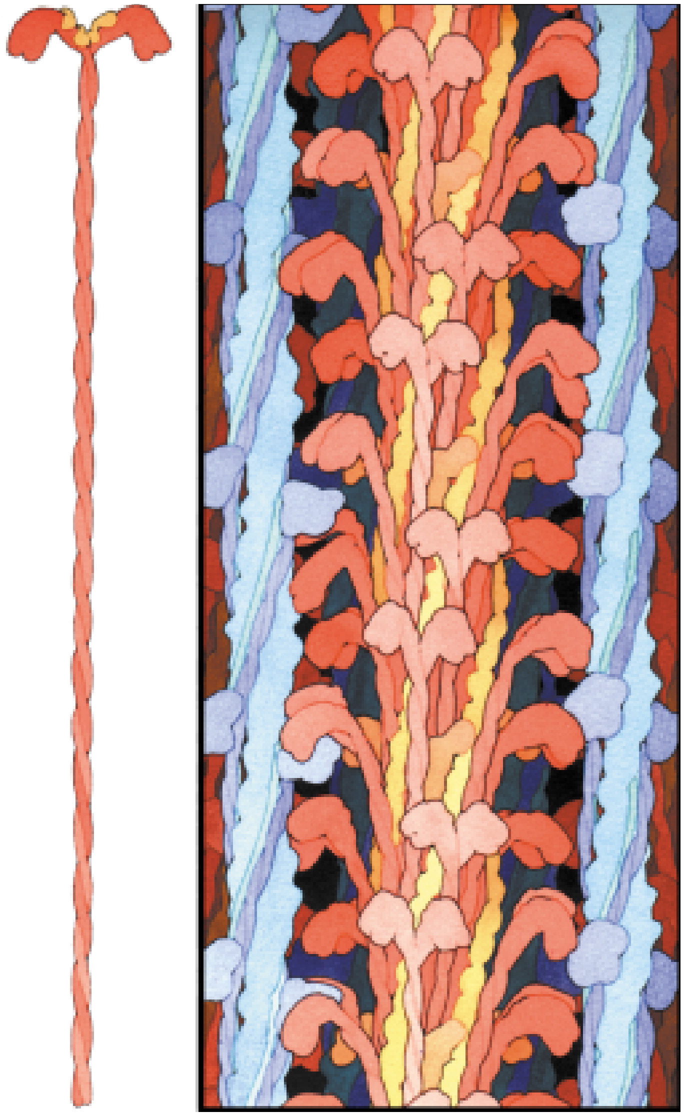
33. (2001): NetWatch. Tools: Reach out and touch. *Science* **292**, p. 1971.

34. K. DeMasters, November 26, 2000: A bank where the currency is molecules. *The New York Times.* **14:3**.

35. M. Nelson (2000): The definitive source for protein structures. *Genome Biology* **1:6**, report 2056.

36. B. Toner (2001): Distributed PDB users can share 3D images using new interactive environment from SDSC. *Bioinform* **5**, pp. 4-6.

37. C. Shirky, October 23, 2000: Seven ways of looking at a protein. *Feed Magazine.* **http://www.feedmag.com/feature/fr409_master.html.**



Myosin is a molecule-sized muscle that uses chemical energy to perform a deliberate motion. The painting above shows how myosin is arranged inside muscle cells. About 300 myosin molecules bind together, with all of the long tails bound tightly together into a large "thick filament." A short segment of a thick filament is shown in red, next to a scale drawing of a single myosin molecule. The many myosin heads extending from the thick filament then reach over to actin filaments, shown in blue and green, and together climb their way up. This illustration by Dr. David Goodsell is part of the Molecule of the Month series.

ABOUT THE COVER

The molecular images featured here were derived from the Molecule of the Month series by Dr. David Goodsell of The Scripps Research Institute. Each month, a key biological molecule is selected for further exploration for a general audience. These images, descriptions about the molecules, and links to related information and structures can be found at **http://www.rcsb.org/pdb/molecules/molecule_list.html.**

Dr. Helen M. Berman, *Director*
Department of Chemistry
Rutgers University
610 Taylor Road
Piscataway, NJ 08854-8087
732-445-4667
Fax: 732-445-4320
berman@rcsb.rutgers.edu

Dr. Phil Bourne, *Co-director*
San Diego Supercomputer Center
University of California, San Diego
9500 Gilman Drive
La Jolla, CA 92093-0537
858-534-8301
Fax: 858-822-0873
bourne@sdsc.edu

Dr. Gary Gilliland, *Co-director*
Biotechnology Division
National Institute of Standards and Technology
Gaithersburg, MD 20899-8310
301-975-2629
Fax: 301-330-3447
gary.gilliland@nist.gov

Dr. John Westbrook, *Co-director*
Department of Chemistry
Rutgers University
610 Taylor Road
Piscataway, NJ 08854-8087
732-445-4290
Fax: 732-445-4320
jwest@rcsb.rutgers.edu

***A list of current RCSB PDB Team members is available at* http://www.rcsb.org/pdb/rcsb-group.html.**

RCSB PARTNERS

Rutgers, The State University of New Jersey

Department of Chemistry
Rutgers University
610 Taylor Road
Piscataway, NJ 08854-8087

National Institute of Standards and Technology

Biotechnology Division
National Institute of Standards and Technology
Gaithersburg, MD 20899-8310

San Diego Supercomputer Center at the University of California, San Diego

SDSC
UC San Diego
9500 Gilman Drive
La Jolla, CA 92093

http://www.pdb.org/

Send questions or comments to:

info@rcsb.org