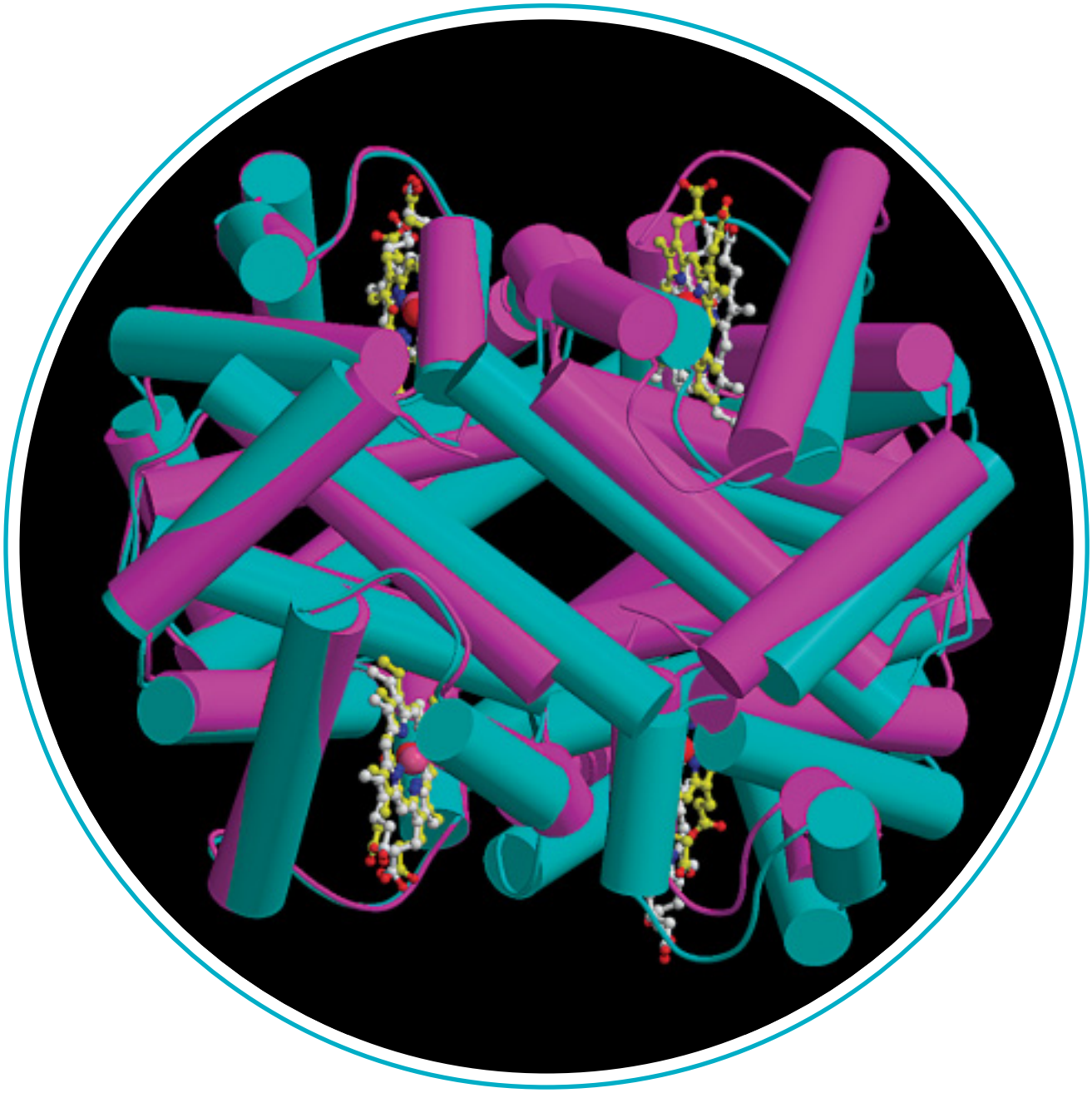
PROTEIN DATA BANK

PDB

ANNUAL REPORT



July 2001 – June 2002

RESEARCH COLLABORATORY FOR STRUCTURAL BIOINFORMATICS

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

CONTENTS

Message from the Director	1
What is the PDB?	2
How Does it Work?	4
Progress & Achievements	6
Recent Publications	10
The Future of the PDB	11

RCSB PARTNER SITES

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

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

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

http://rutgers.rcsb.org/

Center for Advanced Research in Biotechnology of the National Institute of Standards and Technology, Gaithersburg, MD, USA

http://nist.rcsb.org/

DEPOSITION SITES

ADIT

RCSB
http://deposit.pdb.org/adit/

Osaka University
http://pdbdep.protein.osaka-u.ac.jp/adit/

AutoDep

EBI
http://autodep.ebi.ac.uk/

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

The Protein Data Bank (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. The PDB's mission is to provide accurate, well-annotated data in the most timely and efficient way possible to facilitate new discoveries and advances in science.

The RCSB PDB Leadership Team: (left to right) John Westbrook, Philip E. Bourne, Helen M. Berman, and Gary L. Gilliland

This document is the third Annual Report of the Research Collaboratory for Structural Bioinformatics (RCSB) PDB, covering the period from July 1, 2001, through June 30, 2002. It includes many of the noteworthy accomplishments made in the past year. This report on the state of the PDB should serve as a useful guide to the contents and use of the database.

The PDB staff are located at three member institutions of the RSCB: Rutgers, The State University of New Jersey; the San Diego Supercomputer Center (SDSC) at the University of California, San Diego (UCSD); and the Center for Advanced Research in Biotechnology (CARB) of the National Institute of Standards and Technology (NIST). With colleagues and collaborators worldwide, the PDB team has accomplished a great deal in the past year, including:

- Release and consistent update of a complete set of uniform data in mmCIF format.
- Implementation of new query capabilities based on (i) non-redundant sequence data sets and (ii) sequence analysis.
- Distribution of software for data deposition and data validation and software that supports the Corba API for macromolecular structure.
- Progress in the creation of the dictionaries that will be needed for the structural genomics efforts.

Some members of the RCSB PDB Team: (front row, left to right) Anthony Adelakun, Paul Craig, Zukang Feng, David Stoner, Shri Jain, Eliot Clingman, Gary L. Gilliland, Dorothy Kegler, Philip E. Bourne, Nita Deshpande, Gregory Vasquez, David Padilla, Bryan Banister, Li Chen, Huanwang Yang, Lisa Iype, Tania Rose Posa, Kyle Burkhardt, Rose Oughtred, Tammy Battistuz, Christine Zardecki, (back row, left to right) Sharon Cousin, Helge Weissig, Ward Fleri, John Westbrook, Helen M. Berman, Douglas Greer, David Archbell, Gnanesh Patel, Thomas Solomon, Veerasamy Ravichandran, Wolfgang Bluhm, Phoebe Fagan.

In addition, ongoing services continue to be successful, particularly data deposition and annotation, data query, data distribution, and outreach. The international PDB mirror sites provide excellent access, and active help desks allow the PDB staff to be in constant contact with the user community. The foundation provided by these services drives further improvements and refinement of this important resource. Tools are in place to support new challenges. The PDB is prepared for the data from the emerging era of structural genomics, and the PDB development team is constantly developing ways of moving high throughput data in and out of the PDB as efficiently as possible.

The PDB looks to the user community for ways to improve and maintain this resource. We welcome and appreciate your feedback.

Helen M. Berman (Director), Philip E. Bourne, Gary L. Gilliland, and John Westbrook (Co-Directors)
for the Protein Data Bank

WHAT IS THE PDB?

The PDB[1,2] was founded in 1971 at Brookhaven National Laboratory as the international repository for three-dimensional structure data of biological macromolecules. Since July 1, 1999, the PDB has been managed by three member institutions of the RCSB.

The PDB processes, stores, and disseminates structural coordinates and related information about proteins, nucleic acids, and protein-nucleic acid complexes. Some examples of the types of structures of interest that can be found in the PDB archive are enzymes, DNA, RNA, viruses, and ribosomes. The PDB also provides information about related 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 knowledge of the shape of macromolecular structures aids in understanding how these structures function. These structural data assist the pharmaceutical and biotechnology industries in understanding diseases and drug development. 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 biological molecules.

WHAT CAN YOU DO WITH THE PDB?

- Learn what proteins and nucleic acids are and see examples.
- Locate information about a protein that was described in a journal article.
- Create and download pictures of molecules.
- Search for structures with experimental data.
- Explore the sequence and secondary structure of a single molecule or a group of molecules.
- See a structure's entry in related resources and databases.
- Compare a molecule with its structurally similar neighbors.
- Upload and edit data for deposition.
- View and manipulate a structure with a remote collaborator in real time.

New initiatives worldwide focus on structural genomics – high throughput structure determination that is designed to elucidate as many structures as possible of a given proteome, to produce representative structures of all protein families, and to enable a more complete understanding of biochemical pathways. These initiatives will likely produce a great influx of data. New ways to collect, validate, annotate, organize, view, and distribute data are necessary to meet the demands of managing and utilizing such a tremendous amount of information. The PDB will meet this challenge through the incorporation of the most recent technologies available to facilitate the optimal methods for managing structural data.

How is the PDB managed?

The PDB is managed by three member institutions of the RSCB: Rutgers, The State University of New Jersey; the San Diego Supercomputer Center (SDSC), an organized research unit of the University of California, San Diego (UCSD); and the Center for Advanced Research in Biotechnology (CARB) of the National Institute of Standards and Technology (NIST). The RCSB PDB 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. Dr. Helen M. Berman is the Director of the PDB and a Board of Governors Professor of Chemistry and Chemical Biology at Rutgers. 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 Rutgers, which is led by Dr. John Westbrook, Research Associate Professor of Chemistry. Data query and distribution functions are the responsibilities of the PDB team at SDSC/UCSD, which is led by Dr. Philip E. Bourne, Professor of Pharmacology at UCSD and Director of Integrative Biosciences at SDSC. The development and maintenance of the physical archive and CD-ROM production are the responsibilities of the PDB team at CARB/NIST, which is led by Dr. Gary Gilliland, Associate Director of CARB.

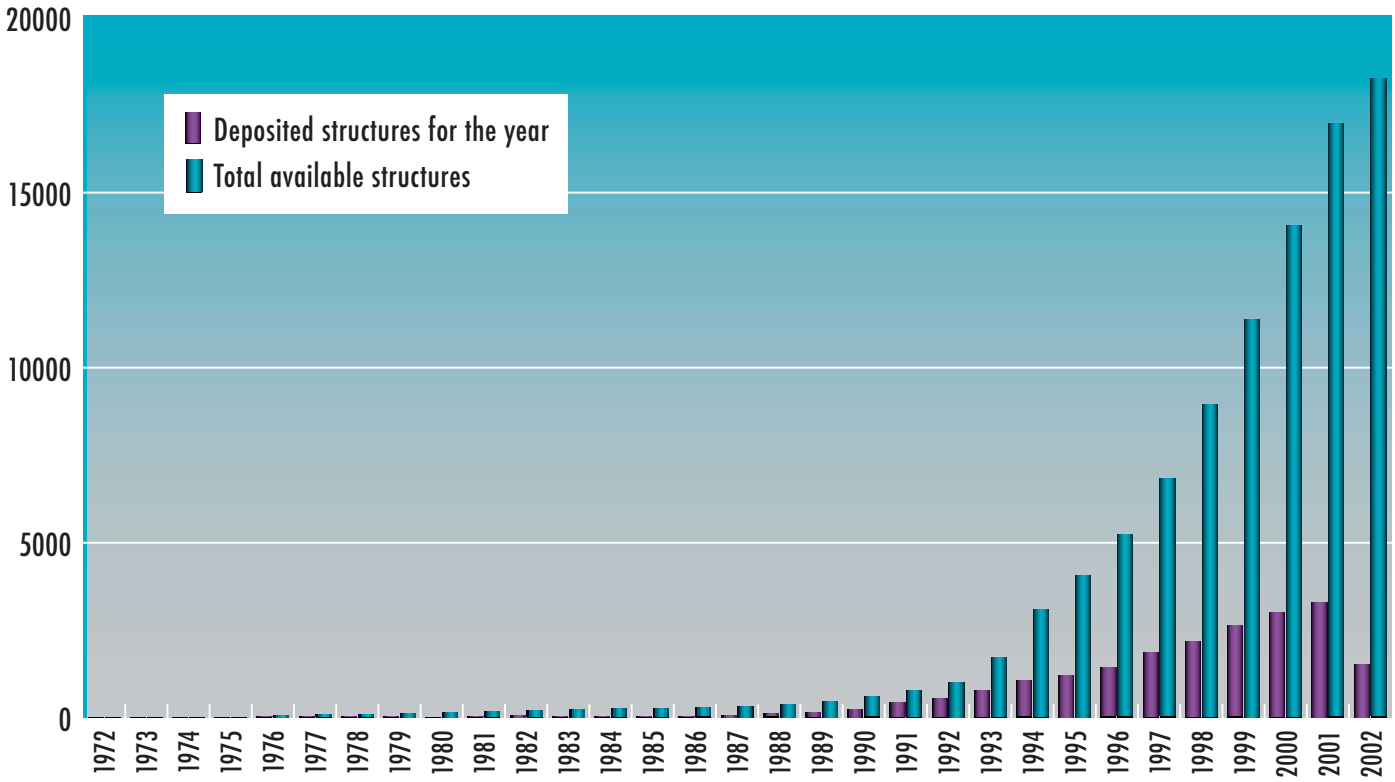
The Mission of the RCSB-PDB Team

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 distributes research results and software. 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. To continue to fulfill this mission, the capabilities of the PDB are being significantly extended.

PDB HOLDINGS AS OF JULY 2, 2002

EXPERIMENTAL TECHNIQUE	PROTEINS, PEPTIDES & VIRUSES	PROTEIN/ NUCLEIC ACID COMPLEXES	NUCLEIC ACIDS	CARBOHYDRATES	TOTAL
X-RAY DIFFRACTION & OTHER	13,998	664	619	14	15,295
NMR	2,276	85	456	4	2,821
TOTAL	16,274	749	1,075	18	18,116

PDB CONTENT GROWTH



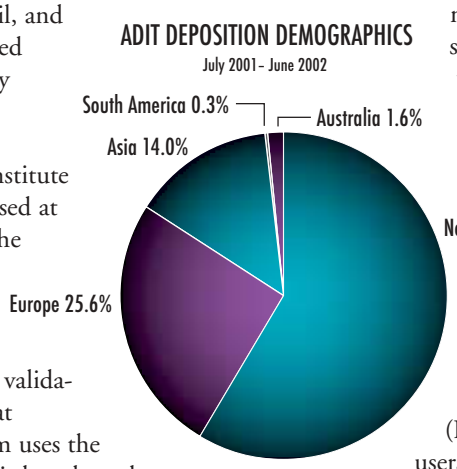
Growth in the number of structures deposited and available in the PDB through June, 2002.

HOW DOES IT WORK?

The PDB is an important biological database. Currently in an average month, the PDB has approximately 290 structures deposited and processed, 200 structures released, and 3.1 million files of individual structure entries downloaded from the archive.

Data Input

A key component of creating the public archive of information is data processing, which entails 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 in the PDB. Data are deposited and processed using the AutoDep Input Tool (ADIT), which is available on-line from sites at RCSB-Rutgers (US) and the Institute for Protein Research (Japan). Data are also accepted via FTP and e-mail, and then processed and annotated using ADIT. Structures may also be deposited using the AutoDep system at the European Bioinformatics Institute (EBI); these data are processed at the EBI and forwarded to the RCSB for release.



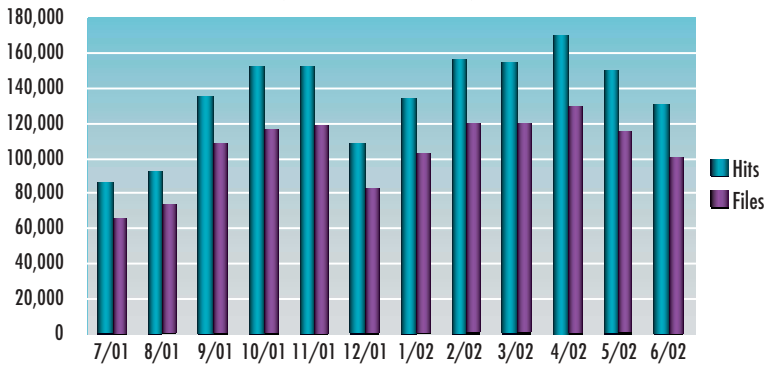
ADIT provides a user interface to a collection of programs for data input, validation, annotation, and format exchange. The ADIT system uses the PDB exchange format that is based on the macromolecular Crystallographic Information File (mmCIF) dictionary (**http://deposit.pdb.org/mmcif/**). mmCIF is an ontology of 1,700 terms defining macromolecular structure and related experiments.[3] 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 available via the Web (**http://deposit.pdb.org/validate/**) and as a software download (**http://deposit.pdb.org/software/**). When finalized, the complete entry, including its status information and PDB ID, is loaded into the core relational database. The PDB staff completes this entire process with an average turnaround of less than two weeks.

Data Distribution and Access

The PDB is a free service available through the Internet. The main PDB Web site at SDSC-UCSD receives an average of more than 130,000 hits per day from all over the world, and on average more than one file is downloaded per second, 24 hours a day, seven days a 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 Gerais (Brazil). A beta Web site is available for users to test new features prior to being incorporated into the main Web site and its mirrors. All RCSB sites are maintained 24 hours a day, seven days a week. New structures are added to the PDB holdings by 1:00 AM Pacific Time each Wednesday, 52 weeks per year.

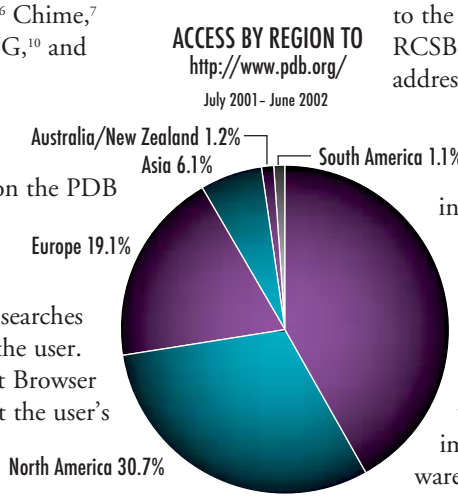
The PDB site offers several different interfaces to query the database. Entering the PDB ID of the target macromolecule in the search box on the home page performs the simplest search. Scientists usually include these IDs when publishing papers describing the structure. A search by PDB ID returns that 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

DAILY AVERAGE QUERY STATISTICS FOR THE PRIMARY RCSB SITE (http://www.pdb.org)



using various methods, are provided along with options to 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 section of the Structure Explorer page. Views of the structure are provided as static images, and in VRML, RasMol,[4] MICE,[5,6] Chime,[7] Swiss-Pdb Viewer,[8,9] STING,[10] and QuickPDB (Java).

Multiple structures can be retrieved by using the keyword search functionality on the PDB 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.



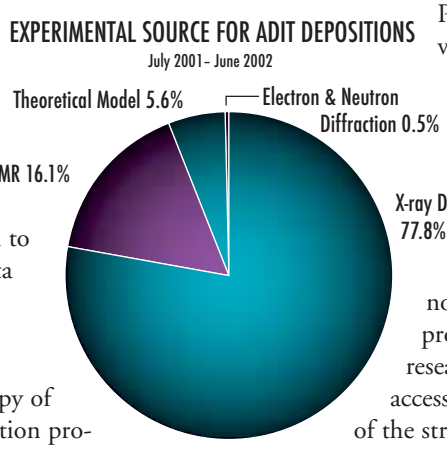
Physical Archive

The PDB Physical Archive contains the files and documents associated with the history of each entry in addition to backup copies of the current data files. This resource, containing paper, magnetic, and electronic records, is maintained at the CARB/NIST site. A back-up copy of the complete query and distribution production system is produced each month by SDSC/UCSD and sent to CARB/NIST for long-term archiving. The overall goal of the PDB Physical Archive is to preserve not only the data submitted by the depositors, but also the records associated with the transactions and activities that are part of the evolution and maintenance of the resource. The access and availability of this information to the PDB staff provides a resource for resolving issues concerning specific entries, aiding in uniformity and value-added annotation, facilitating disaster recovery, and making the information available for research.

Outreach and Education

The PDB facilitates interactions with its user community to provide information about the resource, to gain feedback, and to provide materials that promote a broader understanding of structural biology. This is achieved through accessibility—the PDB maintains an active help desk and has a strong presence at meetings through presentations, user meetings, and exhibit booths. PDB staff members respond, typically within 24 hours, to general and specific inquiries sent to the **info@rcsb.org** electronic help desk. The RCSB-Rutgers site 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 offers a forum for exchanges among members of the PDB community.

The PDB Web site is updated weekly with news, recent developments, and improvements. It includes links to software available for download, mmCIF resources, and educational resources. Dr. David S. Goodsell's Molecule of the Month column, a feature intended for a general audience that focuses on a key biological molecule, is highlighted on the PDB home page each month. The PDB distributes a quarterly newsletter and other online and print materials. PDB activities are also described in various journal papers and articles.



CD-ROM

The PDB distributes a quarterly CD-ROM of its current holdings. Four sets were created and released during the period of this report, at no cost to users. The CD-ROMs are provided by the PDB to help researchers who have limited Internet access or need subsets or a complete set of the structure files for their research.

Advisory Committees

The PDB continues to solicit the advice of several international 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, bioinformatics, and education. A Database Advisory Committee includes six directors of other data resources. The Professional Societies Committee advises on matters related to the PDB's 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 & ACHIEVEMENTS

During the period of this report, development and expansion of the PDB has continued. Efforts in areas such as structural genomics and data uniformity have been productive, ongoing services such as data deposition and query have been enhanced, and new software programs have been released. A collection of highlights from this past year are described below.

Structural Genomics

Structural genomics is a worldwide initiative aimed at determining a large number of protein structures in a high throughput mode. Protein targets range from all open reading frames (ORFs) that encode genes present in the genomes of a variety of organisms, to specific biochemical pathways or to those ORFs associated with specific disease states. In anticipation of an increase in the volume of data from this initiative, the PDB is working to ensure the continuation of timely collection and dissemination of high quality structure data.

The PDB continues to be actively involved in developing the informatics of structural genomics projects. This effort includes active participation in task forces, meetings, and individual interactions with each of the structural genomics centers. The PDB is responsible for maintaining the proposed data deposition specifications and the data dictionary supporting these specifications (**http://deposit.pdb.org/mmcif/**). In addition, the PDB has created software to help integrate data from standard structure determination packages, and established and maintained the Target Registration Database.

The "Structural Genomics Informatics and Software Integration (SG ISI) Workshop" was held May 24-25, 2002 in San Antonio, Texas. Representatives from international structural genomics centers and many key software developers were in attendance. An important outcome of this workshop was the creation of working groups to finalize the specifications for depositing X-ray, NMR, and protein production data generated by structural genomics into the PDB.

The centralized registration database for target sequences from the worldwide structural genomics projects is maintained by the PDB and is available at **http://targetdb.pdb.org/**. The target data can be downloaded as an XML document that follows the recommendations of the International Task Force on Target Tracking (**http://www.nigms.nih.gov/news/meetings/airlie.html**). The target database can be queried by a FASTA sequence search[11] contributing site, protein name, project tracking identifier, date of last modification, or the current status of the target. Search results may be viewed as HTML reports, FASTA data files, or XML documents.

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 PDB entries and maintains a consistent method of annotating current depositions. All PDB files have been rechecked for validity, with errors 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.

During this reporting period, the PDB archive has been standardized and released in mmCIF format for community review at **ftp://beta.rcsb.org/pub/pdb/uniformity/data/mmCIF/**. The files follow the latest version of the PDB exchange data dictionary. This dictionary, available from **http://deposit.pdb.org/mmcif/**, is an extension of the mmCIF dictionary supplement that was developed by the PDB and the EBI. An

IMPORTANT WEB ADDRESSES

PDB Home Page	**http://www.pdb.org/**
PDB FTP Site	**ftp://ftp.rcsb.org/**
Structural Genomics	**http://www.rcsb.org/pdb/strucgen.html**
Software Resources	**http://www.rcsb.org/pdb/software-list.html**
mmCIF Resources	**http://deposit.pdb.org/mmcif/**
Data Uniformity	**http://www.rcsb.org/pdb/uniformity/**
OpenMMS	**http://openmms.sdsc.edu/**
Deposition FAQ	**http://deposit.pdb.org/**
Molecule of the Month	**http://www.rcsb.org/pdb/molecules/molecule_list.html**
Education Resources	**http://www.rcsb.org/pdb/education.html**
PDB Newsletter	**http://www.rcsb.org/pdb/newsletter.html**

application program called CIFTr was made available for translating mmCIF-formatted files into PDB-formatted files. CIFTr works on UNIX platforms, and can be downloaded from http://deposit.pdb.org/software/ (see below for further details).

Deposition and Processing

During the period covered by this report, 3,510 files were deposited to the PDB through an international effort. Data are deposited and processed via ADIT at the RCSB-Rutgers site (US) and at the Institute for Protein Research, Osaka University (Japan), and via AutoDep at the EBI-MSD (UK). These data were usually processed and returned to the depositor in less than two weeks.

Checklists of the data items to have available when depositing structures via ADIT have been made available for both X-ray and NMR depositions. These checklists highlight the information that will be requested when making a deposit.

A variety of programs (described below) were released for data deposition and data processing. These programs are available as open source and executable binary downloads.

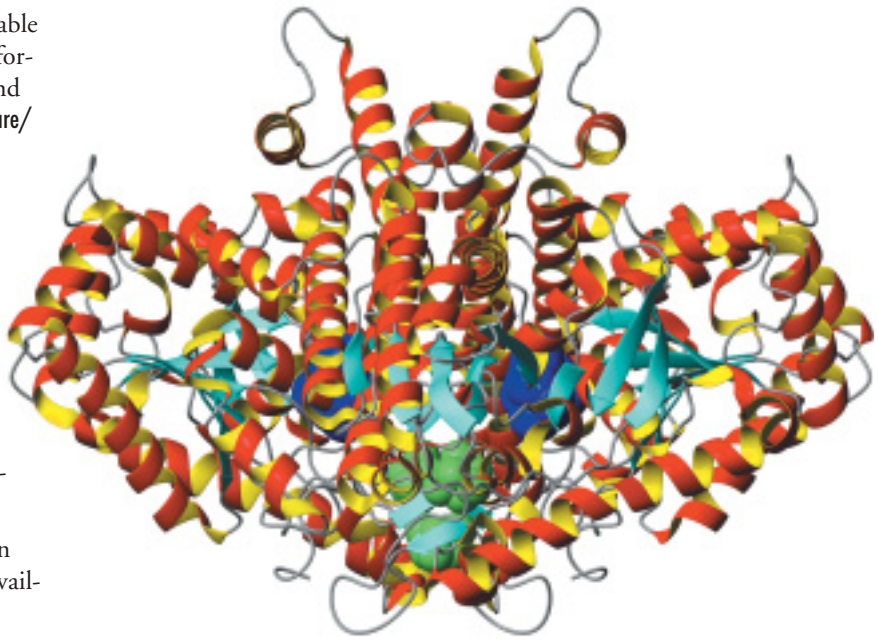
Cryo-Electron Microscopy

In collaboration with the EBI team and the scientific community, all PDB deposition systems have been extended to accept data items specifically describing cryo-electron microscopy (cryo-EM) methods that generate three-dimensional coordinates. This effort has benefited from detailed recommendations on deposition from individual researchers in this field. A prototype of the deposition system was presented at a Gordon Conference on cryo-EM in the summer of 2001. The new versions of ADIT and AutoDep (EBI-MSD) that support these data items for cryo-EM depositions were released in November 2001. Cryo-EM data definitions are available from the PDB mmCIF Web site (http://deposit.pdb.org/mmcif/).

Enhancements to Query, Reporting, and Access

A query of unreleased structures now indicates if the structure factor or constraint file has been deposited, and when the file will be released. This feature was added to the primary PDB Web site and its mirrors after a period of testing on the beta Web site.

Users can now query sequences of yet-to-be-released structures for the purposes of structure prediction and to prevent duplication of structure determination efforts. Through the PDB Status Search interface at http://www.rcsb.org/pdb/status.html, users may query all available sequences, or query based on criteria such as title or deposition date. The prerelease of sequence data in advance of the coordinates is determined solely at the discretion of the depositor, who may also choose to hold the sequence until the structure is released.



In both aerobic and anaerobic microorganisms, the enzyme responsible for carbon monoxide metabolism is carbon monoxide dehydrogenase (CODH). CODH converts CO into CO_2. One of the two principal types of CODHs contains a nickel cluster at its active site. Entry ***1jjy*** *presents the structure of a CODH that contains a [Ni-4Fe-5S] cluster (in dark blue) in each of the homodimeric enzyme's two active sites and three iron sulfur clusters (in green).*

PDB ID: 1jjy

Dobbek, H., Svetlitchnyi, V., Gremer, L., Huber, R., Meyer, O. (2001): Crystal Structure of a Carbon Monoxide Dehydrogenase Reveals a [Ni-4Fe-5S] Cluster. Science ***293****, pp. 1281-5.*

A sequence redundancy filter has been implemented on all PDB Web sites. This additional search option, which is available from all search interfaces, will remove most sequence homologues from the set of structures returned by the user's query. In the Query Result Browser, users are able to toggle between the full set of structures and the reduced set. Sequence clustering uses an algorithm developed in the laboratory of Adam Godzik.[12]

Select components of the STING Millennium Suite (SMS) are now available to PDB users. The Sequence Details and View Structure sections of the Structure Explorer page link to two interactive SMS views for any PDB structure. Users can access both structure and sequence views for a particular PDB entry, which include features such as a graphical display of amino acid contacts. A simpler "Protein Dossier" is also available from the Sequence Details section, offering a static graphical summary of sequence-based properties, such as relative entropy and PROSITE motifs, as well as structure-based properties, such as temperature factors, solvent accessibility, amino acid contacts, and interface (chain contact) regions. The Geometry

section of the Structure Explorer page now links to a Ramachandran plot for each PDB entry, also served from SMS. Finally, hyperlinks found in a structure's Other Sources page have been made available from other Structure Explorer sections for easier accessibility.

As of July 1, 2002, the PDB has separated theoretical model coordinate files from the main archive. Theoretical models are available from the PDB FTP site and continue to be accessible by entering the PDB ID in the search box on the PDB home page.

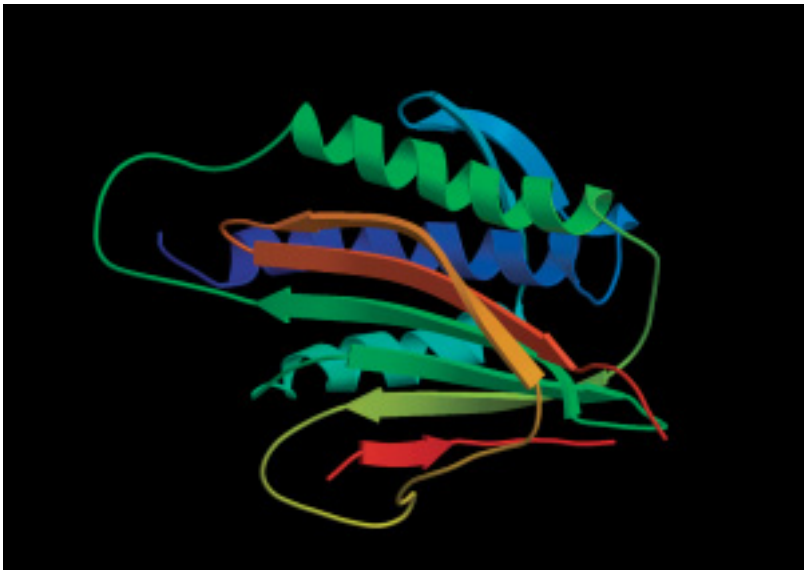
To enhance access to the primary PDB Web site and FTP site, two pairs of load-balanced Enterprise class Sun servers have been procured to administer these sites. The redundant systems ensure access even in the case of a hardware failure. In conjunction with the two independent network paths that now provide Web access to these sites and the increase in Internet bandwidth at the SDSC-PDB site, PDB users now enjoy even more robust connectivity to PDB's resources.

Corba: an Application Programming Interface

The PDB has initiated, developed and promoted the Object Management Group-adopted Corba standard (OMG specification dtc/2001-04-06). Closely aligned with the mmCIF standard, the Corba specification opens the door to seamless and specific access to PDB data by providing a standard applications 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.

A Corba-supporting software suite, the OpenMMS Toolkit, was developed to facilitate the use of macromolecular structure data by various scientific applications. This software contains a set of tools that demonstrate the functionality of the OMG Corba specification. This release is currently in the beta-testing stage. Compiled and source-only distributions of OpenMMS are available at **http://openmms.sdsc.edu/**.

The development of software to support a robust public Corba server is also underway.



Proper cell division during mitosis depends on the even distribution and alignment of chromosomes. Errors in this process can cause aneuploidy, which can result in cancer or birth defects. In order to ensure accurate chromosomal partitioning, spindle checkpoints monitor the attachment of chromosomes to spindle microtubules. The structure in entry ***1klq****, determined by solution NMR, is of a key molecular component of the spindle checkpoint, Mad2 bound to a peptide ligand. This ligand induces a major conformational change in Mad2 that is required for correct checkpoint function.*

PDB ID: 1klq

Luo, X., Tang, Z., Rizo, J., Yu, H., (2002): The Mad2 spindle checkpoint protein undergoes similar major conformational changes upon binding to either Mad1 or Cdc20. Mol.Cell *9(1), pp. 59-71.*

New Software Released

The PDB's Software page is a portal to software developed by the RCSB and others in the macromolecular structure community. RCSB-developed software is available for download, such as the CIFTr application for translating files between mmCIF and PDB formats and software tools for parsing and managing mmCIF files. Links to external software resources relating to mmCIF, crystallography, NMR, structure analysis and verification, modeling and simulation, and molecular graphics are also available from this page.

During this period, additional RCSB-developed programs for mmCIF usage, data processing, and FTP site management have been made available for download. Data validation, deposition, and processing tools are available as open source and binary downloads.

RCSB-Developed Software

- **ADIT** A package for editing and checking structure data entries.
- **bnl2rcsb** A script to convert a BNL-style FTP directory structure to an RCSB-style FTP structure.
- **CIFTr** An application program for translating files in mmCIF format into files in PDB format.
- **MAXIT** An application for the processing and curation of macromolecular structure data.
- **mmCIF loader** An application to load mmCIF data

into relational databases and XML.

- **OpenMMS Toolkit** A suite of Java source code that includes an mmCIF parser, RDBMS loader, XML translator, and Corba server.
- **PDB_EXTRACT** Tools and examples for extracting mmCIF data from structure determination applications.

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). Collaboration with the BMRB, which has now become part of the RCSB, has resulted in a data dictionary for NMR structure and experimental data, and in the creation of a prototype integrated deposition tool.

Outreach and Education

The PDB continues to interact with its user community through a variety of means.

The PDB participated in the exhibitions at the American Crystallographic Association's (ACA) Annual Meeting (May 2002), the Ninth International Conference on Intelligent Systems for Molecular Biology (ISMB, August 2001), and the Biophysical Society Annual Meeting (February 2002). PDB staff members presented talks, demonstrations, and posters at more than thirty meetings around the world, including the 20th European Crystallographic Meeting (August 2001), the Fifteenth Symposium of the Protein Society (August 2001), and the International School of Crystallography Meeting (May 2002). User group meetings were held at the ACA meeting and locally at RCSB sites.

The PDB designed and released a poster entitled "Molecular Machinery: A Tour of the Protein Data Bank." This poster features illustrations of 75 select structures from the PDB, showing their relative sizes at a scale of three million to one, and generally describes their critical roles in the functions of living cells. The content and images were provided by David S. Goodsell. The poster is being distributed to research laboratories and educational institutions around the world. Requests for copies may be emailed to **info@rcsb.org**.

Images from the Protein Data Bank were presented at "The Art of Science," an art exhibit held at The Gallery, a space dedicated to art exhibits at Rutgers University. Various representations of proteins found in the PDB were highlighted, including large-scale depictions of the images available from PDB Structure Explorer pages, images of collagen by Jordi Bella, and pictures from the PDB's Molecule of the Month series.

Visitors at the reception for "The Art of Science" exhibit that displayed images from the Protein Data Bank. The exhibit turned out to be one of the more popular shows held at The Gallery, a space dedicated to art exhibits at Rutgers University.

Four issues of the PDB Newsletter, describing the latest developments of the resource, were published and distributed in print, HTML, PDF and plain text formats during the period covered by this report. Additionally, an archive of 63 legacy newsletters dating from September 1974 to January 1993 were scanned and made available online in PDF format. The history of the Protein Data Bank—the growth of the resource, the means of delivery of the data, and the evolution of standard formats—can be traced through those newsletters.

Collaborations with Other Organizations

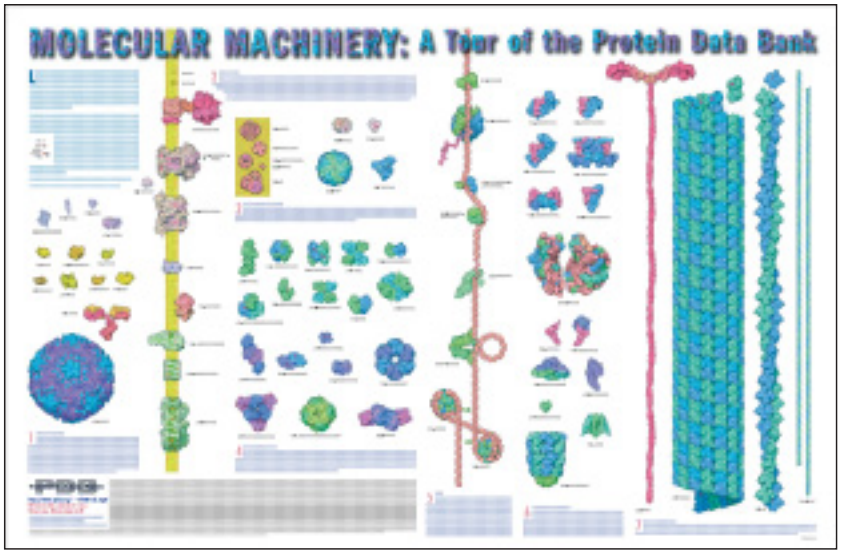
The collaboration on NMR data deposition continues with the BioMagResBank (BMRB), stressing the development of a data dictionary and an integrated deposition system based on ADIT.

The PDB continues to work with the Cambridge Crystallographic Data Centre (CCDC) on methods for ligand validation and now mirrors ReLiBase+, a CCDC ligand resource.

STING Millennium, a Web-based suite of programs that provides simultaneous analysis and display of structure and sequence, is being mirrored through collaboration with the Brazilian Agricultural Research Corporation (Embrapa).

We are working with Emerald Biosystems on crystallization data representation and exchange, and on special handling of structures with licensing restrictions.

The European Bioinformatics Institute (EBI) continues to provide weekly updates of the structures deposited and processed at their AutoDep site. A dic-



The "Molecular Machinery: A Tour of the Protein Data Bank" poster features 75 select structures from the PDB drawn at a relative scale by David S. Goodsell. Requests for copies may be sent to ***info@rcsb.org****.*

tionary that permits all data associated with each deposition to be exchanged has been agreed upon, and the exchange of data based on this dictionary is undergoing testing.

PDB staff members 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 is available at http://srdata.nist.gov/hivdb/.

The PDB is currently working with IBM to evaluate prototype technology for fault-tolerant storage and high-performance database technology.

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. The PDB is also collaborating with this group on an XML representation of PDB data based on the PDB exchange dictionary.

Colleagues at the National Center for Biotechnology Information (NCBI) at NIH continue to work with the PDB on ways to ensure that PDB files can be used by the NCBI-developed databases.

Dr. David S. Goodsell of The Scripps Research Institute continues to contribute the Molecule of the Month feature to the PDB site, which provides introductory level descriptions of the structure and function of key molecules found in the PDB. This collaboration has also resulted in the popular "Molecular Machinery" poster.

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. Eric Martz (University of Massachusetts), Dr. Cherri Pancake (University of Oregon), Dr. Dietmar Schomburg (BRENDA), Dr. Wolf-D. Ihlenfeldt (CACTVS), and Dr. Anthony Williams (ACDLABS).

RECENT PUBLICATIONS

Berman, H.M., Battistuz, T., Bhat, T.N., Bluhm, W.F., Bourne, P.E., Burkhardt, K., Feng, Z., Gilliland, G.L., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J.D., and Zardecki, C. (2002): The Protein Data Bank. *Acta Cryst. D* **58**, pp. 899-907.

Berman, H.M., Goodsell, D.S., and Bourne, P.E. (2002): Protein structures: from famine to feast. *American Scientist* **90**, pp. 350-359.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2001): The Protein Data Bank, 1999 -. In *International Tables for Crystallography*, M.G. Rossman, and E. Arnold, eds. (Dordrecht, Kluwer Academic Publishers), pp. 675-662.

Kuller, A., Fleri, W., Bluhm, W.F., Smith, J.L., Westbrook, J., and Bourne, P.E. (2002): A biologist's guide to synchrotron facilities: the BioSync web resource. *Trends in Biochemical Sciences* **27**: 4, pp. 213-215.

Weissig, H. and Bourne, P.E. (2002): Protein structure resources. *Acta Cryst. D* **58**, pp. 908-915.

Westbrook, J., Feng, Z., Jain, S., Bhat, T.N., Thanki, N., Ravichandran, V., Gilliland, G.L., Bluhm, W., Weissig, H., Greer, D.S., Bourne, P.E., and Berman, H.M. (2002): The Protein Data Bank: Unifying the archive. *Nucleic Acids Research* **30**, pp. 245-248.

THE FUTURE OF THE PDB

The PDB has many notable achievements. PDB staff members will continue to improve 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

The PDB will continue to streamline deposition and data processing procedures. This will be especially important when the structural genomics projects start to generate large amounts of new structures. The PDB will continue to strengthen the data processing software that has been released as source code and as executables.

Structural Genomics

Community collaborations to develop the required data items for structural genomics will continue. The PDB will also work with the software developers and beamline operators to integrate the output of applications programs with PDB deposition software. Efforts to produce tools to facilitate the extraction and integration of the data from existing structure determination software will be expanded. The PDB will further encourage the structure genomics centers to use PDB software tools in their respective data processing operations.

A new deposition interface will be provided for the structural genomics initiatives. This interface will include the expanded set of data items recommended for deposition by the International Task Force on the Deposition, Archiving, and Curation of the Primary Information.

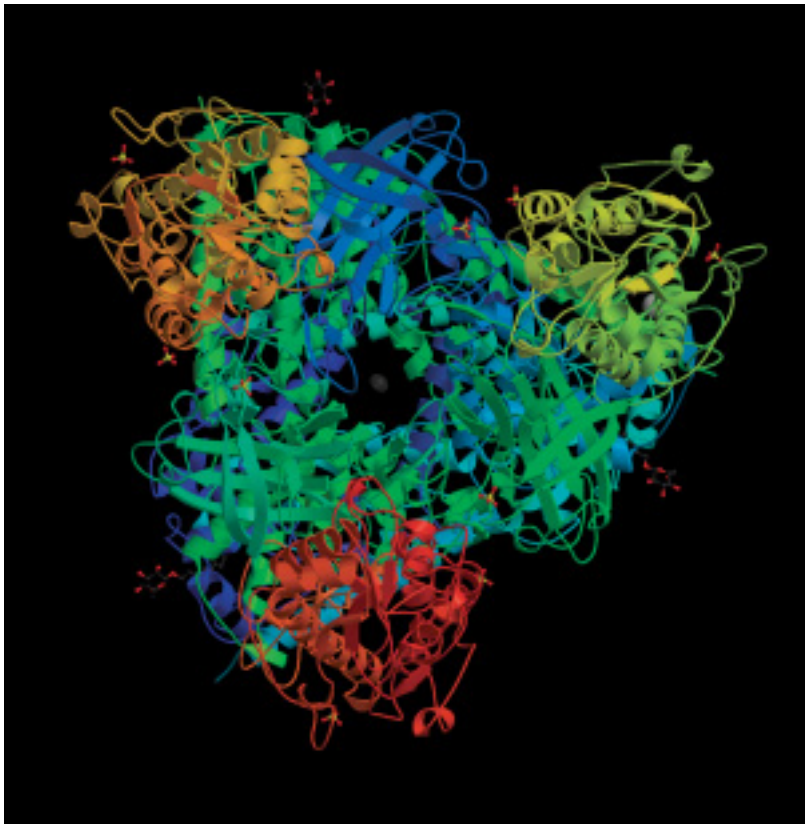
Data Distribution and Query

During the past year, the PDB has undertaken a complete redesign of its query and distribution system to improve its future extensibility and maintenance. The redesign, which is currently in a pre-alpha phase of development, is supported by a relational database management system and Java v2 Enterprise Edition (J2EE), and will offer new capabilities for browsing and searching the PDB archive while maintaining all the currently available functionalities. Usability and navigation of the Web site will be greatly improved, with an enhanced help system and site search option. This effort makes use of the data from the Data Uniformity Project, and will demonstrate the remediation improvements, offering even more accurate query results.

Establishment of the new PDB mirror Web site at the Max-Delbrück Center for Molecular Medicine (MDC) in Berlin, Germany is approaching completion, and this site will become available in the near future.

NMR

The PDB will continue to maintain an active dialog with the NMR community through contacts with individual depositors, structural genomics centers, the NMR Taskforce, and as an active participant in the



Reoviruses are double stranded RNA viruses that infect both plants and animals. Reoviruses are icosohedral in shape and unlike other icosohedral viruses, are not covered by a phospholipid envelope that can be used to fuse with the outer membrane of the host cell. As a result, reoviruses must penetrate the host cell in a different way. The majority of the outer protein coat of the virus is composed of two proteins, $\mu 1$ and $\sigma 3$. During infection, $\sigma 3$ is proteolytically cleaved off, exposing $\mu 1$, which is responsible for membrane penetration. The structure presented in entry ***1jmu****, determined by X-ray crystallography, is of a heterohexameric complex ($\mu 1_3$ $\sigma 3_3$) of these two proteins that reveals several aspects of the penetration mechanism.*

PDB ID: 1jmu

Liemann, S., Chandran, K., Baker, T.S., Nibert, M.L., Harrison, S.C. (2002): Structure of the Reovirus Membrane-Penetration Protein, $\mu 1$, in a Complex with its Protector Protein, $\sigma 3$. Cell ***108****, pp. 283-95.*

CCPN software initiative. The development of a common interface for the coordinates, constraints, and chemical shift data with the BMRB is progressing. The completion of the NMRIF dictionary and common Web interface is anticipated within the next project year.

Data Uniformity

Efforts on the PDB Data Uniformity Project will proceed on an ongoing basis. The PDB will continue to ensure that legacy and current files are annotated consistently. Synonym lists will be expanded and maintained.

CD-ROM

CD-ROM production will continue on a quarterly schedule. Beginning with the Fall 2002 release, experimental data – structure factors and constraint files – will be made available as a CD-ROM set separate from the coordinate files. This will decrease the number of disks sent to subscribers.

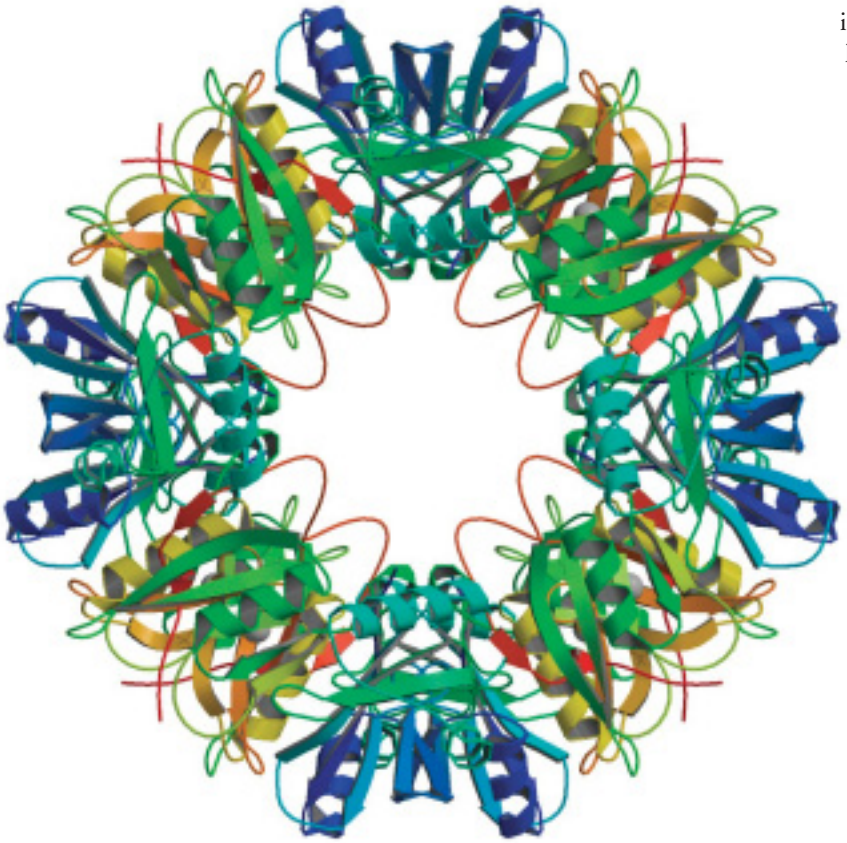
Physical Archive

With the reorganization of the physical archive materials, the next phase of working with the physical archive will involve scanning and electronically storing documents associated with the PDB. A diverse set of paper files associated with depositions were scanned and are under review to establish which of the documents should be made electronically available for the PDB staff's use. The files include correspondences and administrative records about individual entries.

The recovery of electronic files from the legacy magnetic media is continuing. To accommodate the restored data, one terabyte of disk storage has been installed on the computer systems being used to house the electronic archive. As data are restored, they are incorporated into a file structure that is designed to trace their origin, that aids in retrieving information, and that can be easily maintained and expanded. Efforts are underway to establish a unique (non-redundant) file set.

Outreach and Education

The PDB will continue to solicit the input of its diverse user community of researchers, teachers, and students through interactions at future meetings and through help desk services. Publications and online information resources will continue to develop, while novel ways to educate and convey the PDB's mission to new audiences will be explored. The PDB will remain engaged with users to ensure that this resource continues to serve as a prominent resource in the evolving era of structural genomics.



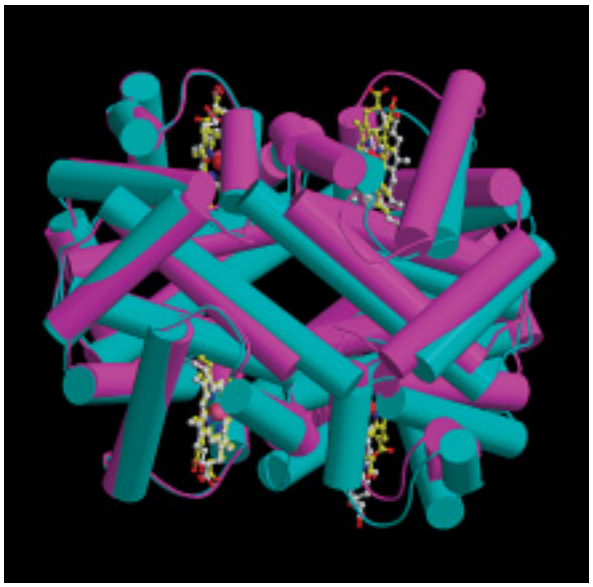
2,3-dihydroxybiphenyl-1,2-dioxygenase (DHBD) functions within the biphenyl biodegradation pathway of several micro-organisms. Entry ***1kmy*** *presents the structure of DHBD complexed with its substrate 2,3-dihydroxybiphenyl.*

PDB ID: 1kmy

Vaillancourt, F.H., Han, S., Fortin, P.D., Bolin, J.T., Eltis, L.D. (1998): Molecular Basis for the Stabilization and Inhibition of 2,3-Dihydroxybiphenyl 1,2-Dioxygenase by t*-Butanol.* J. Biol. Chem. ***273****, pp. 34887-95.*

Vaillancourt, F.H., Barbosa, C.J., Spiro, T.G., Bolin, J.T., Blades, M.W., Turner, R.F., Eltis, L.D. (2002): Definitive evidence for monoanionic binding of 2,3-dihydroxybiphenyl to 2,3-dihydroxybiphenyl 1,2-dioxygenase from UV resonance Raman spectroscopy, UV/Vis absorption spectroscopy, and crystallography. J. Am. Chem. Soc. ***124****(11), pp. 2485-96.*

ABOUT THE COVER



Two horse hemoglobin structures are overlaid to show the conformational change that occurs when a ligand is bound to the heme group. The pioneering work of Dr. Max Perutz in X-ray crystallography of proteins won him the Nobel Prize in 1962. His life-long work on the allosteric mechanism for the cooperative binding of oxygen to hemoglobin won him respect throughout the scientific community. We wish to acknowledge the enormous debt we owe to Dr. Perutz by featuring his structures on this annual report.

PDB ID: 2mhb

Ladner, R.C., Heidner, E.J., Perutz, M.F. (1977): The structure of horse methaemoglobin at 2.0 Å resolution. J. Mol. Biol. *114, pp. 385-414.*

PDB ID: 2dhb

Bolton, W., Perutz, M.F. (1970): Three-dimensional Fourier synthesis of horse deoxyhaemoglobin at 2.8 Ångstrom units resolution. Nature *228, pp. 551-2.*

SELECTED REFERENCES

1. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E. (2000): The Protein Data Bank. *Nucleic Acids Res.* **28**, pp. 235-242.
2. Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, E.E., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., Tasumi, M. (1977): Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, pp. 535-542.
3. Bourne, P.E., Berman, H.M., Watenpaugh, K., Westbrook, J.D., Fitzgerald, P.M.D. (1997): The macromolecular Crystallographic Information File (mmCIF). *Meth. Enzymol.* **277**, pp. 571-590.
4. Sayle, R., Milner-White, E.J. (1995): RasMol: biomolecular graphics for all. *Trends Biochem. Sci.* **20**, p. 374.
5. Bourne, P.E., Gribskov, M., Johnson, G., Moreland, J., Weissig, H. (1998): *Pacific Symposium on Biocomputing*, pp. 118-129.
6. Tate, J.G., Moreland, J., Bourne, P.E. (1999): MSG (Molecular Scene Generator): a Web-based application for the visualization of macromolecular structures. *Journal of Applied Crystallography* **32**, pp. 1027-1028.
7. MDL Information Systems Inc. Chime. United States.
8. Guex, N., Peitsch, M.C. (1997): SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. *Electrophoresis* **18**, pp. 2714-2723.
9. Kaplan, W., Littlejohn, T. (2001): Swiss-PDB Viewer (Deep View). *Brief Bioinform* **2**, pp. 195-197.
10. Neshich, G., Togawa, R., Vilella, W., Honig, B. (1998): STING (Sequence To and withIN Graphics) PDB_Viewer. *Protein Data Bank Quarterly Newsletter* **85**, pp. 6-7.
11. Pearson, W.R., Lipman, D.J. (1988): Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U.S.A.* **24**, pp. 2444-2448.
12. Li, W., Jaroszewski, L., Godzik, A. (2001): Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* **17**, pp. 282-283.

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. Gary L. 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. Philip E. 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. 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.

As part of the purine salvage pathway, 5´-Deoxy-5´-methylthioadenosine phosphorylase (MTAP) catalyzes the phosphorolysis of 5´-deoxy-5´-methylthioadenosine (MTA) to adenine and 5-methylthio-D-ribose-1-phosphate. The crystallographically determined structure of MTAP presented in entry ***1jdt*** *contains 6 identical subunits, each of which contains one active site.*

PDB ID: 1jdt

Appleby, T.C., Mathews, I.I., Porcelli, M., Cacciapuoti, G., Ealick, S.E. (2001): Three-Dimensional Structure of a Hyperthermophilic 5´-Deoxy-5´-Methylthioadenosine Phosphorylase from Sulfolobus solfataricus. J. Biol. Chem. ***276****, pp. 39232-42.*

RCSB PARTNERS

Rutgers, The State University of New Jersey

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

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

SDSC
UC San Diego
9500 Gilman Drive
La Jolla, CA 92093-0537

Center for Advanced Research in Biotechnology of the National Institute of Standards and Technology

Biotechnology Division and Informatics Data Center
100 Bureau Drive
Gaithersburg, MD 20899-8314

http://www.pdb.org/

Send questions or comments to:

info@rcsb.org