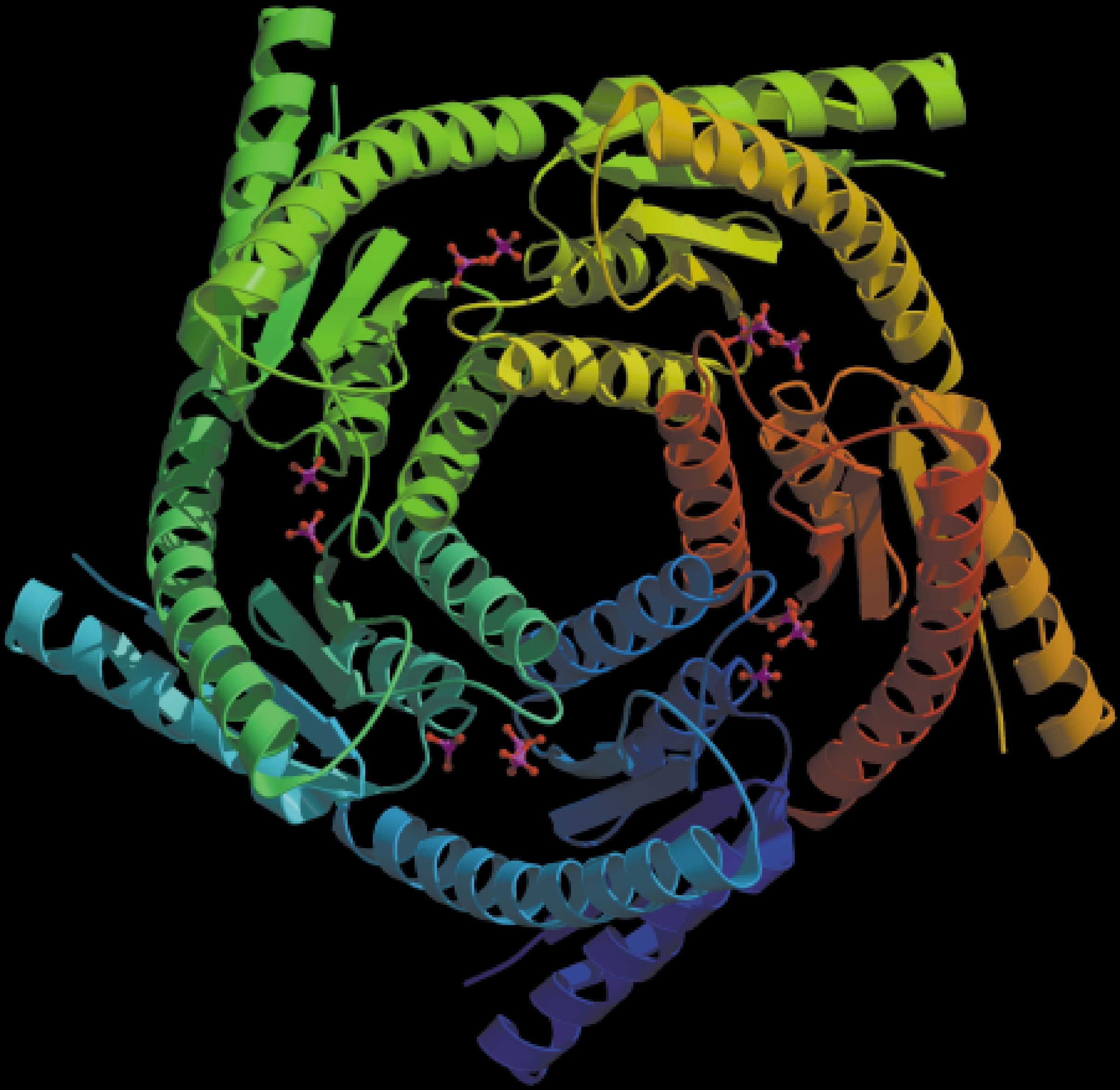
PDB PROTEIN DATA BANK

ANNUAL REPORT

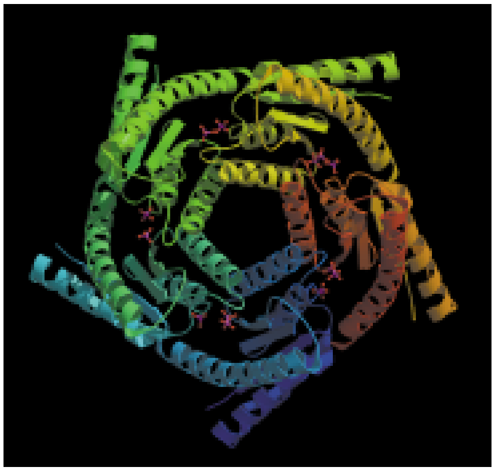
JULY 1999-JUNE 2000



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



ABOUT THE COVER:

Lumazine synthase catalyzes one of the final steps in the synthesis of riboflavin in plants, fungi, and microorganisms. This type of enzyme displays two quaternary structures, the polyhedral protein coat in plants and bacteria, and the beautiful pentameric forms in yeast and fungi as shown here. This molecule was derived from Brucella abortus, *the infectious organism of the disease brucellosis in animals. The active sites of the two structure types are virtually identical, indicating that inhibitors to these enzymes could be effective pharmaceuticals across a broad species range.*

PDBid: 1DIO

B.C. Braden, C.A. Velikovsky, A.A. Cauerhff, I. Polikarpov, F.A. Goldbaum (2000): Divergence in macromolecular assembly: X-ray crystallographic structure analysis of lumazine synthase from Brucella Abortus. J.Mol.Biol. ***297****, p. 1031.*

Message from the Director

Welcome to the first annual report for the Protein Data Bank (PDB) following its transition to the management of the Research Collaboratory for Structural Bioinformatics (RCSB). The period of time between July 1, 1999 and June 30, 2000 represents the first full year of management of the PDB by the RCSB. In this year, the PDB has achieved significant progress in enhancing and maintaining the PDB. This report, prepared for a general audience, briefly describes the purpose and functions of the PDB, details our accomplishments during this first year of RCSB management, and touches upon our plans for the coming year.

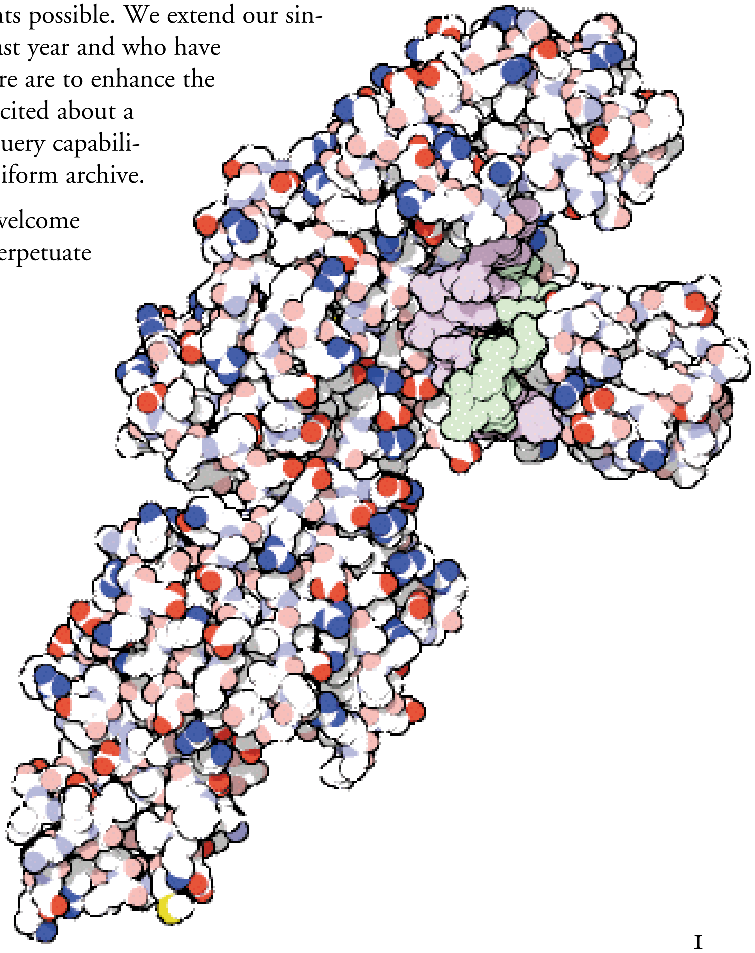
As the sole international repository for three-dimensional structure data of biological macromolecules, the PDB is an important resource for the extensive life-science research community. The RCSB's mission for the PDB is to enable new science. Each of the RCSB partner sites contributes to the operation and development of the PDB: Rutgers, data deposition and processing; National Institute of Standards and Technology (NIST), data uniformity, exploring issues specific to nuclear magnetic resonance (NMR), and data archiving; and the San Diego Supercomputer Center (SDSC) at the University of California, San Diego (UCSD), data query reporting and distribution. We have been able to make great strides in enhancing the PDB due to the unique environment the RCSB offers in personnel, hardware, software, and network infrastructure. Some of the successes that highlight this year of the PDB's management by the RCSB are detailed in this document. Here are just a few of the PDB's outstanding achievements from this timeframe:

- The seamless transition of the PDB from Brookhaven National Laboratory was completed a full three months ahead of schedule
- A large number of files have been processed with a rapid turnaround time
- Legacy data has been reprocessed and cross-referenced to ensure reliability
- An average of 90,000 hits per day have been accommodated by the main PDB Web site alone

The input from a number of collaborators made these achievements possible. We extend our sincerest gratitude to all who have shown their support during the past year and who have helped to make the PDB a better resource. Our plans for the future are to enhance the PDB's many features with new capabilities. We are particularly excited about a higher, faster throughput of deposited data; a greater number of query capabilities, including more complex and accurate queries; and a more uniform archive.

We look forward to working with you in the years to come, and welcome your input to help maintain the integrity of the resource and to perpetuate its importance as a vital enabling technology for the sciences.

Helen M. Berman
on behalf of the entire project team



Taq DNA polymerase has revolutionized biotechnology. This molecule is used in a process which allows small samples of DNA to be duplicated and analyzed. DNA polymerase duplicates genetic information by taking a single strand of DNA and building a complimentary strand. In this image, a small piece of DNA is bound with the template strand in purple and the new strand in green. This image was created from the PDB file 1tau by David S. Goodsell of The Scripps Research Institute.

PDBid: 1TAU

S.H. Eom, J. Wang, T.A. Steitz (1996): Structure of Taq polymerase with DNA at the polymerase active site. Nature ***382****, p. 278.*

What is the PDB?

The Protein Data Bank is the sole international repository for 3-dimensional structure data of biological macromolecules. Specifically, it is a resource that processes, stores, and disseminates structure coordinates and related information about proteins, nucleic acids, nucleic acid complexes, viruses, polypeptides, and some carbohydrates.

Why is it important?

The 3-D structures of proteins and other biological macromolecules contained in the PDB hold significant promise for the pharmaceutical and biotechnology industries in the search for new drugs with few or no side effects, and the efforts to understand the mystery of human disease. Medical researchers envision gaining new insights into the causes, effects, and treatment of various diseases by unlocking the therapeutic potential of biological macromolecules. This requires very precise and accurate information about the atomic structure of complex molecules. The understanding of what a structure looks like aids in understanding how it works.

The PDB provides researchers with a rich source of information about biological structures. Because of the improvements that the RCSB has introduced, PDB users can now access new services and formulate complex queries that will provide reliable answers to further the research efforts of the international biological and biomedical communities. The PDB now provides powerful tools to help researchers understand biological function through investigation of sequence and molecular structure. The tremendous influx of data that is being fueled by the structural genomics initiative, and the increased recognition of the value of structural data in understanding biological function, demand new ways to collect, organize, and distribute data. The PDB will continue to meet this demand using the most modern technology that facilitates the use and analysis of structural data and that creates an enabling resource for biological research.

Who is involved in running it?

The PDB is managed by the Research Collaboratory for Structural Bioinformatics. The RCSB is a non-profit consortium composed of 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 at San Diego (UCSD). The RCSB is supported by funds from the National Science Foundation, the Department of Energy, and two units of the National Institutes of Health: the National Institute of General Medical Sciences and the National Library of Medicine.

The RCSB Project Team manages the overall operation of the PDB. Director Helen M. Berman, Professor of Chemistry at Rutgers, was part of the original team that developed the PDB in 1971 and is the founder of the Nucleic Acid Database. Data deposition and processing are the responsibilities of the RCSB Team at Rutgers, which is led by John Westbrook, Research Associate Professor of Chemistry. Data uniformity, NMR, and the master archive are the responsibilities of the RCSB Team at NIST, which is led by Gary Gilliland, Chief of the Biotechnology Division of the Chemical Science and Technology Laboratory. Data query and distribution functions are the responsibility of the RCSB Team at SDSC, which is led by Phil Bourne, Professor of Pharmacology at UCSD and Senior Principal Scientist at SDSC, and Peter Arzberger, Executive Director of SDSC.

Mission

The mission of the RCSB is to enable science worldwide by providing resources to improve our understanding of structure-function relationships in biological systems. The RCSB integrates a variety of production-level data and software resources, and shares research results and software. The RCSB is dedicated to fostering new scientific advances by providing accurate, consistent, well-annotated 3-D structure data that is delivered in a timely and efficient way to a wide audience. The RCSB will continue to significantly extend the capabilities of the PDB.

Historical background

The PDB was established at the Brookhaven National Laboratory (BNL) in 1971, initially holding 7 structures (Bernstein et al. 1977). After 27 years, responsibility for the operation and enhancement of the Protein Data Bank transitioned from BNL to the RCSB during the period from October 1998 to June 1999. Since July 1, 1999, the RCSB has had sole responsibility for the management of the PDB (Berman et al. 2000).

PDB Holdings (27-Jun-2000)

Experimental Technique	Molecule Type				
	Proteins, Peptides, and Viruses	Protein/Nucleic Acid Complexes	Nucleic Acids	Carbohydrates	Total
X-ray Diffraction and Other	9320	467	515	14	10316
NMR	1605	65	321	4	1995
Theoretical Modeling	246	18	17	0	281
Total	11171	550	853	18	12592

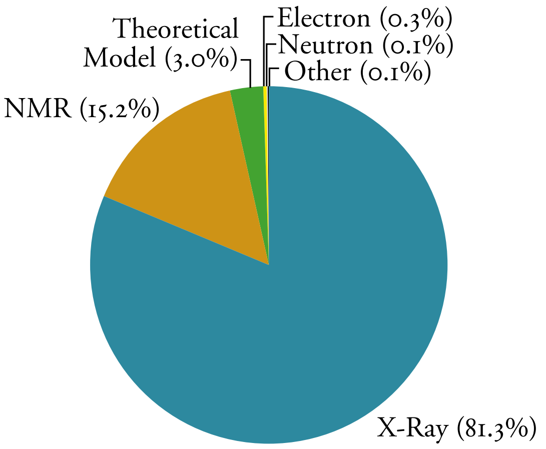
How does it work?

As in any active database, new structures are constantly being added, existing entries are being refined and "cleaned up", users are continually accessing data, and the entire database is periodically archived.

Data input

A key component of creating the public archive of information is the efficient capture and curation of data – data processing. This consists of data deposition, validation, and annotation. The RCSB Team at Rutgers is responsible for all aspects of primary data processing. Data are primarily deposited to the PDB and processed using the AutoDep Input Tool (ADIT). Data are also accepted via ftp, e-mail, CD-ROM, and the legacy BNL AutoDep system. ADIT is built on top of the mmCIF dictionary (Bourne 1997), which is an ontology of 1700 terms defining macromolecular structure and the experiment used to determine the structure. After checks are performed by the PDB staff, validation reports and a completed PDB file are returned to the depositor for review. The complete entry, including its status information and PDB ID, is loaded into the core relational database. This process is completed by the PDB staff with an average turnaround of about two weeks. Data can also be deposited using ADIT at Osaka University (Japan), and by using AutoDep at the European Bioinformatics Institute (United Kingdom). Data are processed at these sites and forwarded to the RCSB for release.

Experimental Source for RCSB Depositions (7/99–6/00)



From July 1999 to June 2000, 2693 files were deposited with the PDB. The 2693 deposited structures come from different experimental sources and from researchers on five continents, as the figures demonstrate (see page 4).

Data retrieval and storage

The PDB is a free service available to the general public via the Internet. The main PDB Web site is maintained by the RCSB Team at SDSC, and receives an average of 90,000 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, located at the RCSB partner sites at NIST and Rutgers, and at Osaka University in Japan, the National University of Singapore, the Cambridge Crystallographic Data Centre (CCDC) in the United Kingdom, and the Universidade Federal de Minas Gerias in Brazil. A beta test site also provides PDB users with the opportunity to try new features before they are incorporated into the main site. These sites are maintained 24 hours a day, 7 days a week. New structures are added to the PDB holdings each Wednesday by 1:00AM Pacific Time.

For users of macromolecular structure data, PDB is a portal to general information about single structures, substructures, and their interrelationships. The query capabilities include access to information across a broad range of structures, a feature vital for comparative analysis. The site offers several different interfaces that can be used to query the database. The simplest search is performed by entering the PDB ID of the desired molecule, which returns a single structure via the Structure Explorer page. The Structure Explorer presents options to further study the molecule, such as the secondary structure or sequence information. Each Structure Explorer entry provides links to the atomic coordinates, crystallization conditions, derived geometric data, structure factors (where available), 3-D structure neighbors computed using various methods, and hyperlinks to other resources which include analysis programs and the primary MEDLINE entry. Dynamic links to the structure's entry in other databases as created by the Molecular Information Agent (MIA; see **http://mia.sdsc.edu**) are available under Other Sources. Many options are available for displaying structures, including VRML, RasMol (Sayle, Milner-White 1995), Chime®, and QuickPDB (Java). Multiple structures can be retrieved by using either the SearchLite interface, which performs a simple keyword search, or the SearchFields interface, which is a more advanced, customizable search based on parameters that the user selects. 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 directly from the PDB ftp server.

Master Archive

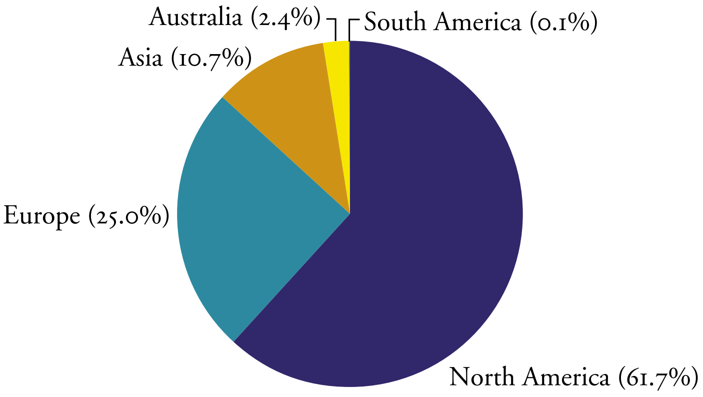
The Master Archive, containing both paper and electronic records, is maintained at the RCSB-NIST site. The paper records are being converted to electronic form for long-term, automatic, reliable access. The electronic storage of all records will ensure the long-term viability of the resource. A snapshot of the complete query and distribution production system is made by SDSC and sent to NIST for long term archiving each month. Currently, depositor tapes that were stored at BNL are being converted to CD-ROM format for reference uses and long-term storage. The PDB also maintains and distributes a quarterly CD-ROM snapshot of its holdings for users who may not have facile internet access or who prefer to have a local copy of the PDB files.

Data Uniformity

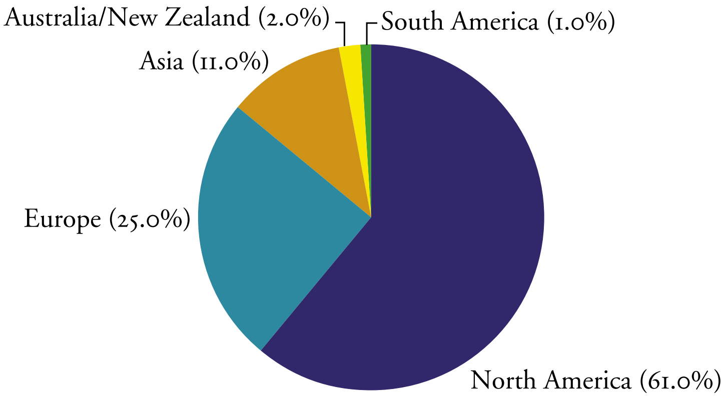
A significant objective of the PDB is to make the archive as consistent and error-free as possible. Query across the complete PDB has been limited by missing, erroneous, and inconsistently reported experimental data, nomenclature, and functional annotation. The evolution of experimental methods, functional knowledge of proteins, and methods used to process these data over the years has introduced various inconsistencies into the PDB archive and has inspired different versions of the PDB format. In addition to concentrating on making current depositions to the PDB consistent, efforts are being made as part of the Data Uniformity Project to enhance the consistency of existing entries.

The RCSB Team at NIST is responsible for the data quality enhancement of the legacy data. The goal of these efforts is to provide a higher level of query capability through the higher curation of the database. The legacy data are being checked for accuracy, uniformity, and completeness in both individual and global respects. Much progress has been achieved in this "clean-up" process by means of two complementary methods used to update and unify the archive. *File-by-file* uniformity processing brings each entry up to the current PDB format by adding records that were not present in some early entries, correcting unresolved problems, providing standard nomenclature, and evaluating data using current validation software. Approximately 3000 entries have been processed in this manner. Additionally, key records within each PDB entry have been targeted for *archive-wide* uniformity processing wherein data are examined and updated on certain global parameters such as synonyms, common names, and names used by other data centers to enrich the reliability of query results.

RCSB Deposition Demographics*

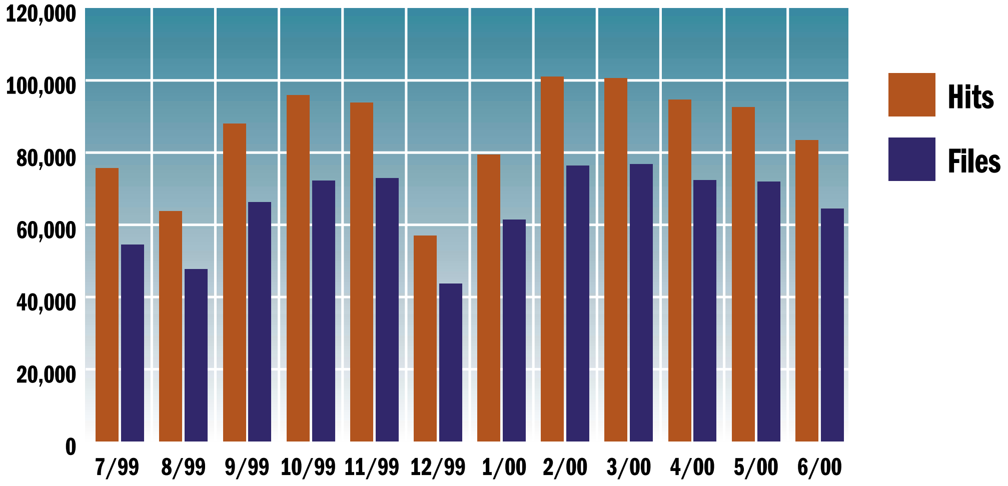


PDB Hits for the Primary RCSB Site (www.rcsb.org) Based on Region (781,857 total)*



**From 7/99 to 6/00*

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



How well is the RCSB doing?

In its first full year of operating the PDB, the RCSB has achieved many noteworthy successes. The PDB Advisory Committee (PDBAC), which consists of members of the international scientific community, commended the RCSB on its high level of service and highlighted a number of RCSB achievements during its site visit in March, 2000:

- Smooth, seamless transition three months ahead of schedule
- Timely and efficient processing of newly deposited and backlogged information
- Development of a new and robust data input tool (ADIT)
- Introduction of the mmCIF data format, which allows greater flexibility for incorporating new forms of data
- "Cleaning up" older PDB entries to ensure data uniformity
- Proactive investigation of XML and CORBA software technologies
- Enhancement of database query tools that meet the needs of most users
- Improved handling of Nuclear Magnetic Resonance structures
- Continued collaboration with the European Bioinformatics Institute (EBI) data deposition site
- New collaboration with the Osaka University (Japan) ADIT data deposition site
- Education and outreach efforts to the scientific user community
- Ongoing efforts to collaborate with other bioinformatics databases and organizations

The following text elaborates on many of these and describes other accomplishments.

The RCSB Collaboration

One of our primary achievements is the RCSB itself. Rutgers, UCSD, and NIST have successfully worked together to maintain and improve the PDB to meet the growing demands of the scientific community. The participants at all three locations work together in order to achieve the shared vision of providing the best possible bioinformatics database to the international scientific community. While each RCSB site has primary responsibilities, members of each of the sites contribute to aspects of the project at the other sites. This is necessary to achieve the integrated systems that we have developed and utilized. The Project Leadership Team consists of Dr. Helen Berman and Dr. John Westbrook of Rutgers, Dr. Phil Bourne and Dr. Peter Arzberger of UCSD, and Dr. Gary Gilliland and Dr. T.N. Bhat of NIST. They, and all RCSB members, communicate frequently by phone and e-mail to coordinate activities and track progress.

Advisory Committees

The PDB has several Advisory Committees who provide advice and guidance. The members of the PDB Advisory Committee were selected for their combined expertise in X-ray crystallography, NMR, modeling, and bioinformatics. The RCSB has also established a Database Advisory Committee whose members consist of directors of other international data resources. This committee helps with matters specific to databases and has already helped the RCSB to arrive at a plan for releasing remediated data. The Professional Societies Committee will provide advice on matters relevant to our interactions with professional societies. The NMR Task Force provides guidance with respect to meeting the specific needs of the NMR community.

Large number of entries processed per year

During the summer and early fall of 1999 the RCSB completed the processing of the 468 unreleased structures inherited at the time of the PDB transition. The RCSB also reprocessed 456 structures that BNL had released as "Layer 1" entries, meaning unannotated and unprocessed, and brought them to the completely processed "Layer 2" status. In addition, the RCSB processed and released approximately 2300 newly deposited structures. In all, the RCSB processed and released approximately 25% of the entries in the entire PDB in its first full year of operation. As a result of the talented and highly motivated annotation staff at Rutgers and the use of ADIT software, the RCSB has been able to fully process files at the time of deposition, with an average turnaround time of less than two weeks.

Large number of hits/users

Over the past year, the primary PDB Web site, located at SDSC, has averaged over 90,000 hits a day, with a peak average of over 100,000. The adjacent figure shows the daily average of hits and files downloaded by month. The staff at the SDSC is dedicated to maintaining the site on a 24/7 basis and has succeeded with only a few hours of down time this past year.

Outreach and Education

The RCSB is dedicated to providing a better understanding of biological macromolecules. To accomplish this, the PDB has implemented a number of educational and outreach efforts to address the needs of a very diverse user community ranging from students to distinguished professionals.

Molecule of the Month

One such innovation is the popular Molecule of the Month piece, which was highlighted in the journal *Science* (2000). Each month a key biological molecule is selected for further exploration. Beautiful images of the molecule are provided by David Goodsell of the Scripps Research Institute and featured on the PDB home page. Subsequent links provide additional information about the structure and function of the molecule at a general level.

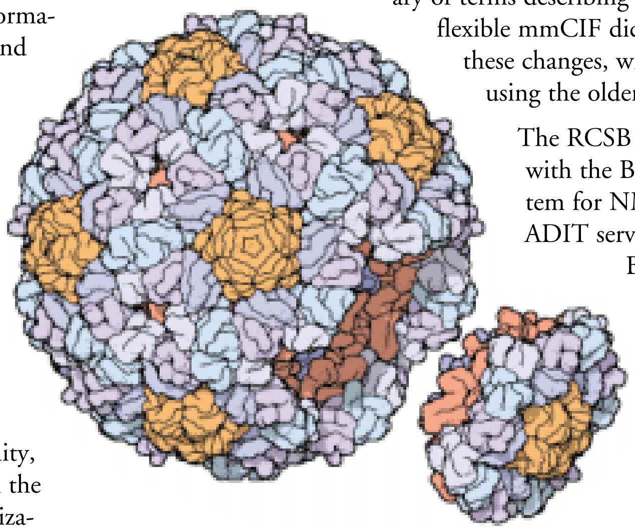
Get Educated

The site's "Get Educated" page includes an introduction to proteins for general audiences and materials for undergraduates on topics such as nucleic acids and nucleic acid-containing structures, principles of protein structure, and electron microscopy. Several beginners' tutorials on how to query the PDB and how to use RasMol and the Swiss-PDB Viewer (Guex, Peitsch 1997), two popular molecular graphics viewing programs, are available. Links are frequently added to this resource, which also includes papers on the PDB, animated presentations about the PDB, and VRML "protein documentaries" developed by students.

Ask the PDB Community

Another important service is the electronic help desk at **info@rcsb.org**, which is available to answer all types of questions about the PDB, usually within a 24-hour period. The RCSB-Rutgers site also maintains two other addresses: **deposit@rcsb.rutgers.edu**, for general deposition and processing questions; and **help@rcsb.rutgers.edu**, for ADIT information. Furthermore, the **pdb-l@rcsb.org** discussion list provides a forum for users to interact and collaborate.

The PDB staff is also actively involved at conferences, hosting exhibit booths, demonstrations, and user group meetings to gather feedback from the user community and to provide information about PDB's capabilities and growth. Materials such as informative flyers and a quarterly newsletter are circulated to a broad audience to inform the community about PDB's resources and new items of interest.



A bacteriophage is a virus that attacks bacteria. The phiX174 *bacteriophage, pictured here, attacks the common human bacteria* Escherichia coli, *infecting the cell and forcing it to make new viruses. This molecule was the 10,000th deposited structure, a milestone for the PDB.*

PDBid: 1CD3

T. Dokland, R.A. Bernal, A. Burch, S. Pletnev, B.A. Fane, M.G. Rossmann (1999): The role of scaffolding proteins in the assembly of the small, single-stranded DNA virus phiX174. J. Mol. Biol. ***288****, p. 595.*

Collaboration with other non-RCSB organizations

In order to better serve the needs of the scientific community, the RCSB is collaborating with the following bioinformatics organizations:

BioMagResBank (BMRB): The RCSB is coordinating the deposition of NMR data with this site. The BMRB is responsible for experimental data and the PDB is responsible for structural data.

Cambridge Crystallographic Data Centre (CCDC): The CCDC and the RCSB have begun to develop methods for ligand validation. The RCSB is also working to mirror ReLiBase, which is now under the management of the CCDC.

European Bioinformatics Institute (EBI): The EBI provides weekly updates of the structures deposited at their site.

Institute for Protein Research, Osaka University: Structures deposited at the ADIT site at Osaka University are processed and then forwarded to the RCSB PDB for release.

National Center for Biotechnology Information (NCBI): We are working with NCBI to ensure that our files can be used by the databases developed and distributed by NCBI.

NCI-Frederick Cancer Research and Development Foundation: We are working to move the HIV Protease Database from this foundation to NIST and incorporate PDB data files that have undergone the PDB uniformity process.

Swiss Institute for Bioinformatics/Glaxo: Both parties are working to create an mmCIF based dictionary that can be used to store model data.

NMR Collaborations

The PDB's continuing NMR efforts have involved working closely with the NMR community to develop a data dictionary and deposition/validation tools and procedures specific for NMR. This is being accomplished by working directly with the BMRB, by forming a PDB NMR Task Force, and by ensuring a presence at all major NMR meetings and workshops. Achievements to date include improved handling of ensembles of NMR structures and the development of an expanding dictionary of terms describing NMR structure determinations. The flexible mmCIF dictionary has permitted implementation of these changes, without compromising searches on files using the older nomenclature.

The RCSB has undertaken a collaborative effort with the BMRB to develop a joint deposition system for NMR-specific data based on ADIT. An ADIT server for the data items collected by the BMRB has been built and is being alpha-tested. This version uses an mmCIF-like dictionary that was derived from the NMR STAR deposition form used by BMRB. The goal is to provide a single integrated deposition system for NMR data that will accept both experimental and structure data.

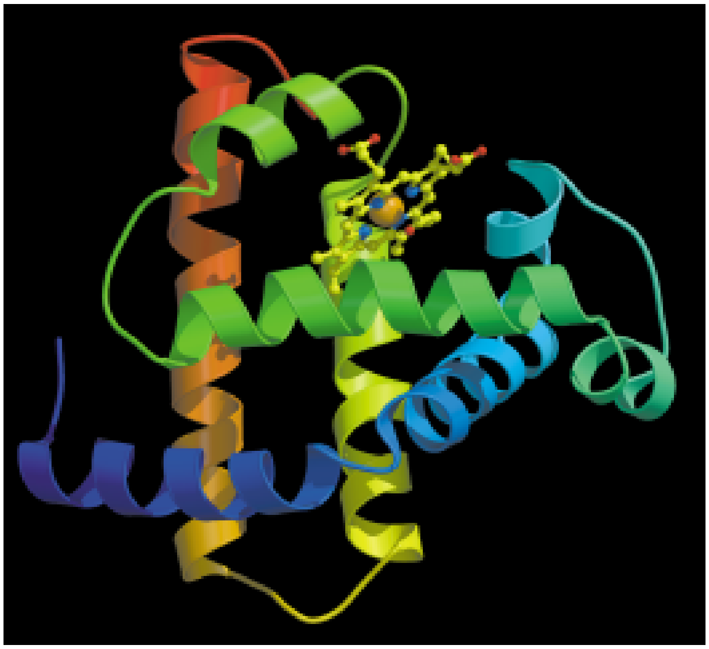
The PDB is also participating in the Collaborative Computing Project for NMR (CCPN) with the goal of developing data-harvesting capabilities. Data harvesting refers to the exchange of data between the software packages used during the determination of a structure and the deposition software at the public databanks. Much of the information required for deposition is required by the software packages used in processing experimental data and determining structures. Automatic exchange of these data will ease the burden on the depositor and provide richer and more accurate data to the databanks with improved data uniformity. This project requires development of a generally accepted data format, accompanied by a standard library of tools for manipulating data conforming to this standard.

Where is the PDB going?

Future challenges and plans to meet them

Expected growth of data

These are exciting and challenging times to be responsible for the collection, curation, and distribution of macromolecular structure data. It is expected that improved technologies and experimental methods will accelerate the growth of the number of structures in the coming years. It is estimated that the PDB could grow to approximately 35,000 structures by 2005, which would nearly triple its size in less than five years. One major factor in this growth is the advent of structural genomics, whose goal is to determine the structures of as many of the proteins accessible from a given genome as possible, in the shortest time possible. This promises to greatly increase the amount of information that needs to be archived within the PDB, presenting a huge challenge to the timely distribution of high quality data. The PDB's approach of using modern data management practices should permit it to scale to accommodate a large data influx. Structural genomics programs will be capturing a substantial body of interim results and biophysical data that are likely to be of interest to the scientific community. The RCSB is uniquely positioned to address the latter challenge because of its expertise with mmCIF and modern database technologies.



The science of protein structure began with myoglobin. After years of arduous work, John Kendrew (Myoglobin and the Structure of Proteins, Nobel Lecture, December 11, 1962) and his coworkers determined the atomic structure of myoglobin, laying the foundation for an era of biological understanding. Myoglobin is a small, bright red protein. It is very common in muscle cells, and gives meat much of its red color. Its job is to store oxygen, for use when muscles are hard at work.

PDBid: 1MBN

H.C. Watson (1969): The Stereochemistry of the protein myoglobin. Prog.Stereochem. *4, p. 299.*

Expected growth in access

As technology advances, it is anticipated that the PDB's user base will expand with a greater need for ease of access. In order to accommodate this demand, the RCSB plans to enhance the robustness of the PDB's query capabilities. The implementation of synonyms for proteins and ligands used in the PDB for querying is one way of enhancing these capabilities. One notable feature to be released in the future is myPDB. This service will allow users to register a query that will be run automatically at a specified frequency and return the results via email. New non-redundant data sets will be available soon along with the traditional non-redundant set known as PDBselect, originally derived by Hobohm, Sander et al. (1992). Property based searching features will also be made available, along with new display features such as Ramachandran plots, packing diagrams, and the complete biological unit.

The RCSB will also continue its efforts to develop a standard application interface for macromolecular data based on the Common Object Request Broker Architecture (CORBA). An initial proposal was submitted to the Object Management Group (OMG) in February 2000. When the final specification is accepted, the PDB will be able to publish a robust and efficient interface definition for direct use by programs and other databases accessing the PDB. Standards will be agreed upon and implemented.

Planned hardware upgrades to the current system include the addition of two 4-processor Sun 450's each with 6 GB of memory and 200 GB of disk space to permit expansion and faster access to the PDB primary production site. The existing 4000's will be used for the beta site and the proposed myPDB resource.

Another access development pertains to the data archive. The RCSB is proceeding with the next phase of archiving the physical data, which involves scanning and electronically storing all documents associated with the PDB. Protocols will be developed and implemented regarding procedures for handling the old and new documents. The CD-ROM production will continue on a quarterly schedule, and during the coming year the development of a DVD distribution product will be pursued to reduce the number of CDs in the set and their production costs.

Future of Data Processing

The on-going maintenance and development of data processing and data validation software will continue to be a primary software effort. With the anticipated increase in the number of structures submitted to the PDB for deposition, greater automation of the tasks associated with annotation and validation will remain a high priority. The protocols established for dealing with the legacy data will help to provide mechanisms for updating the PDB in the future, as notations and usage patterns change. The PDB must also accommodate new requirements, such as those that will come with depositions of cryogenic electron microscopy (cryo-EM) data.

The RCSB plans to release ADIT and its validation tools, and will provide support. The RCSB will base its future user interface development plans for ADIT on the feedback that it obtains from the first distribution of ADIT.

Addressing the data integration tasks associated with the new structural genomics projects will be a major focus for next year. The RCSB plans to establish a liaison with each of the structural genomics projects to serve as a point of contact for data exchange issues. The PDB could serve as a model for many data representation and management systems that will be required to support the structural genomics effort. The RCSB will promote the use of the PDB tools and methods among the structural genomics projects.

The RCSB will also continue to work with software developers to promote the use of mmCIF in structure determination software, and will continue to work within the CCPN project to promote the use of a data representation that will integrate well with mmCIF.

Developments in Data Uniformity

Data uniformity work will continue by focusing on structure classification, compound records, chain ID fields, refinement parameters, coordinates, sequence records, and the biological unit. All the information generated as database tables during the uniformity process will be accessible from the PDB query and report functions, and will be stored in mmCIF files separate from the original archive and made available to all users of the PDB, along with new tools for accessing these mmCIF files.

Future of NMR

The PDB will continue to maintain an active dialog with the NMR community through its NMR Task Force, working closely with the BMRB, and as an active participant in the CCPN software initiative.

Outreach and Education activities

The RCSB is committed to maintaining a proactive attitude in further developing its outreach to all sectors of the scientific and nonscientific community. The RCSB will continue to expand the resources in place, such as the Get Educated resource, while developing new features. The PDB will continue to be represented at a variety of meetings and conferences.

Efforts are also under way to further improve awareness and usability of PDB data at different levels of expertise and use. New methodologies for the integration of traditional database information and functional content are being developed using eXtensible Markup Language (XML), a technology receiving a great deal of attention by Web developers.

The RCSB will continue to promote frequent interactions with the community and will rise to meet the needs of its growing user base.

Selected References

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

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**, pp. 235-242.

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

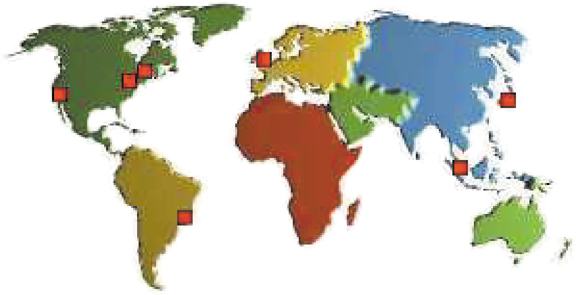
R. Sayle, E.J. Milner-White (1995): RasMol: Biomolecular graphics for all. *Trends in Biochemical Sciences* **20** (9), p. 374.

Chime Pro® is a registered trademark of MDL Information Systems, Inc. in the United States.

Net Watch HOT PICKS (2000): Molecular profiles. *Science* **287**, p. 1883.

N. Guex, M.C. Peitsch (1997): SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. *Electrophoresis* **18**, pp. 2714-2723.

U. Hobohm, M. Scharf, R. Schneider, C. Sander (1992): Selection of representative protein data sets. *Protein Science* **1** (3), pp. 409-417.



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/

This molecule is the first complete atomic structure of the 50S ribosomal subunit, which catalyzes peptide bond formation; it binds initiation, termination, and elongation factors. The model includes 2711 of the 2923 nucleotides of 23S ribosomal RNA, all 122 nucleotides of its 5S ribosomal RNA, as well as structures for the 27 of the 31 proteins in the subunit. This is a unique deposition as it is the first comprehensive structure of its kind to be included in the PDB holdings.

PDBid: 1FFK

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

RCSB PROJECT TEAM LEADERS

Dr. Helen M. Berman
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
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
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
Department of Chemistry
Rutgers University
610 Taylor Road
Piscataway, NJ 08854-8087
732-445-4290
Fax: 732-445-4320
jwest@rcsb.rutgers.edu

Dr. Peter Arzberger
San Diego Supercomputer Center
University of California, San Diego
9500 Gilman Drive
La Jolla, CA 92093-0505
858-534-5079
Fax: 858-822-0948
parzberg@sdsc.edu

Dr. T.N. Bhat
Biotechnology Division
National Institute of Standards and Technology
Gaithersburg, MD 20899-8310
301-975-8702
Fax: 301-975-8717
bhat@nist.gov

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

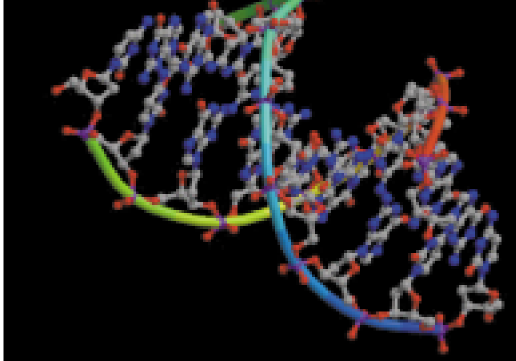
San Diego Supercomputer Center at the University of California, San Diego

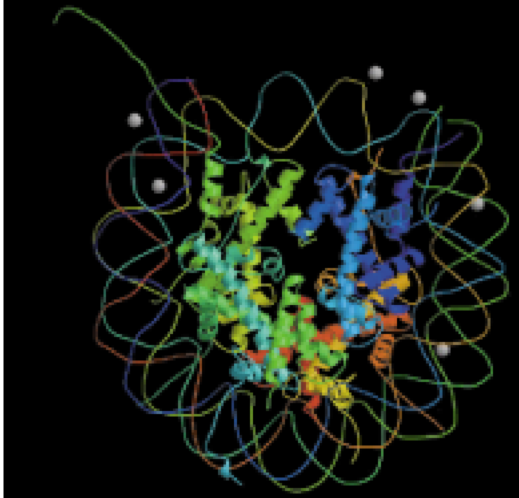
SDSC
UC San Diego
9500 Gilman Drive
La Jolla, CA 92093

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

Send questions or comments to:

info@rcsb.org





TOP: X-ray crystallography can reveal the three-dimensional structure of short fragments of DNA with amazing precision. It provides information on the overall helical structure and the configuration of local features such as base-pair stacking patterns and backbone form, which also gives such clues to its function as potential binding behavior. The groove geometries and accessibilities of this self-complementary DNA fragment CCG-GCGCCGG may be important for the potential binding of both proteins and drug molecules to G/C stretches in DNA.

PDBid: 1CGC

U. Heinemann, C. Alings, M. Bansal (1992): Double helix conformation, groove dimensions and ligand binding potential of a G/C stretch in B-DNA. EMBO. J. ***11****, p. 1931.*

BOTTOM: The structure of the nucleosome core particle of chromatin shows in detail how the histone protein is assembled into eight similar sections called "octamers", and how 146 base pairs of DNA are organized into a superhelix around it. This molecule reveals the form of DNA that is predominant in living cells and offers a wealth of information on DNA binding and bending by the histone octamer.

PDBid: 1AOI

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 A resolution. Nature ***389****, p. 251.*