RCSB PDB PROTEIN DATA BANK

ANNUAL REPORT
for July 2004 – June 2005

RESEARCH COLLABORATORY FOR STRUCTURAL BIOINFORMATICS:
Rutgers, The State University of New Jersey
San Diego Supercomputer Center, University of California, San Diego
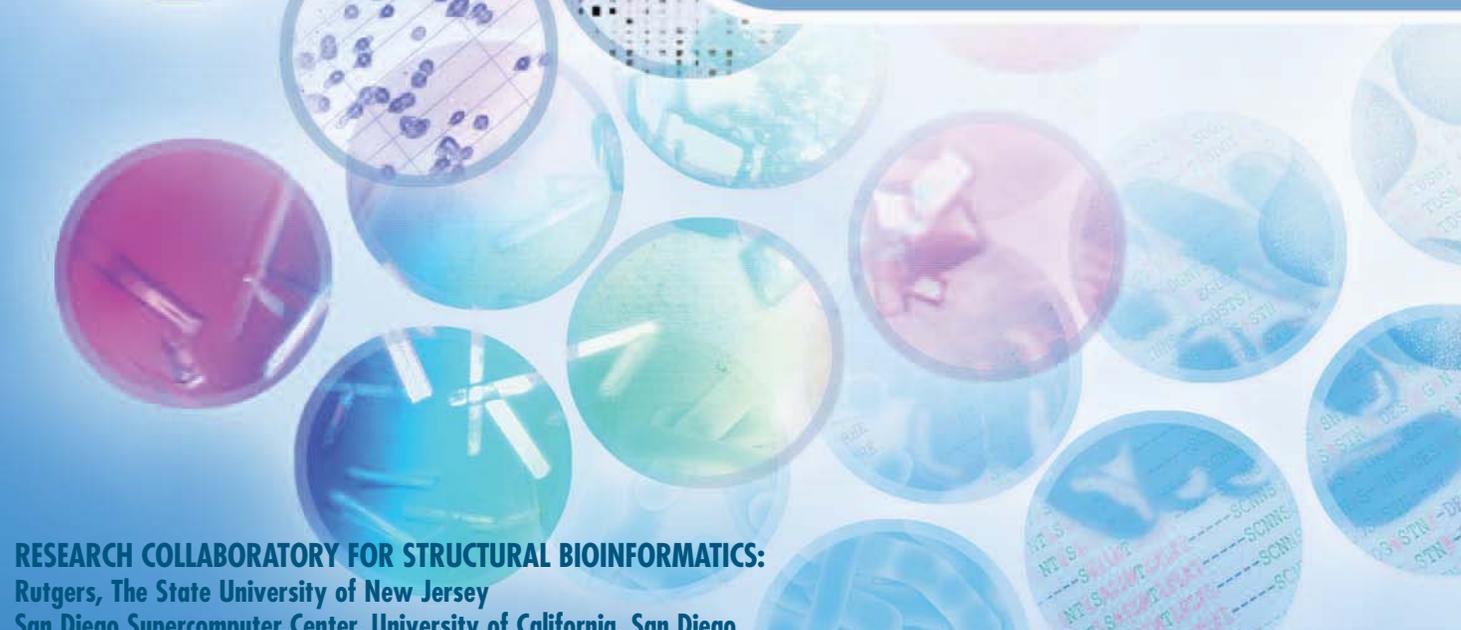
About the Cover

The data that the RCSB PDB tracks from structural genomics projects – from target selection to structure determination – are illustrated here. The structures shown are from the nine pilot centers participating in the first phase of the NIH-funded Protein Structure Initiative.

1t3u
(interior)
Unknown conserved bacterial protein from *Pseudomonas aeruginosa* PAO1
New York Structural Genomics Research Consortium

1nkt
Crystal structure of the SecA protein translocation ATPase[4]
Mycobacterium tuberculosis (TB) Structural Genomics Consortium
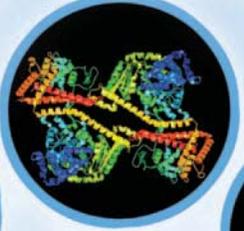
1rjj
Solution structure of a homodimeric hypothetical protein from *Arabidopsis thaliana*[1]
Center for Eukaryotic Structural Genomics

1nc7
Crystal structure of *Thermotoga maritima* 1070
Midwest Center for Structural Genomics

1sq4
Crystal structure of the putative glyoxylate induced protein from *Pseudomonas aeruginosa*
Northeast Structural Genomics Consortium

1xn4
Putative Mar1 ribonuclease from *Leishmania major*
Structural Genomics of Pathogenic Protozoa Consortium
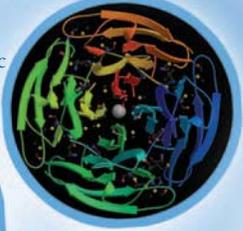
1n0e
Crystal structure of a cell division and cell wall biosynthesis protein UPF0040 from *Mycoplasma pneumoniae*[5]
Berkeley Structural Genomics Center

1pry
Structure determination of fibrillarin homologue from hyperthermophilic archaeon *Pyrococcus furiosus*[2]
Southeast Collaboratory for Structural Genomics
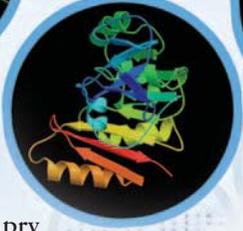
1o4v
Crystal structure of phospho-ribosylaminoimidazole mutase PurE (Tm0446) from *Thermotoga maritima*[3]
Joint Center for Structural Genomics

Contents

Message from the Director

The RCSB PDB is uniquely positioned to track the growth of structural biology. In terms of shear volume, the numbers are impressive. More than one hundred structures are deposited every week, each with its own story to tell. Some come from small laboratories that focus on a single aspect of biology aimed at cracking the mystery of how proteins function. Others come from structural genomics groups where high throughput methods make it possible to rapidly determine structures on a genomic scale and give us an even broader view of the many possible shapes of proteins. Now we can see structures of large macromolecular machines that have been determined using a combination of powerful methods, including cryo-electron microscopy.

The RCSB PDB is up to the challenge of handling the data produced by all of these initiatives with methods for managing both its volume and complexity. It is both daunting and exciting to be part of this era in which the secrets of biology are being revealed at the molecular level.

The RCSB is not alone in managing the data from these new initiatives. Through the worldwide PDB (wwPDB) we are working with our partners in Japan (the PDBj) and Europe (MSD-EBI) to ensure that the data remain uniform and freely available.

This report presents the many different RCSB PDB activities in data deposition, data access and education. We hope you find that it reflects the excitement we feel about our contribution to the burgeoning field of structural biology.

Helen M. Berman
Director, RCSB PDB
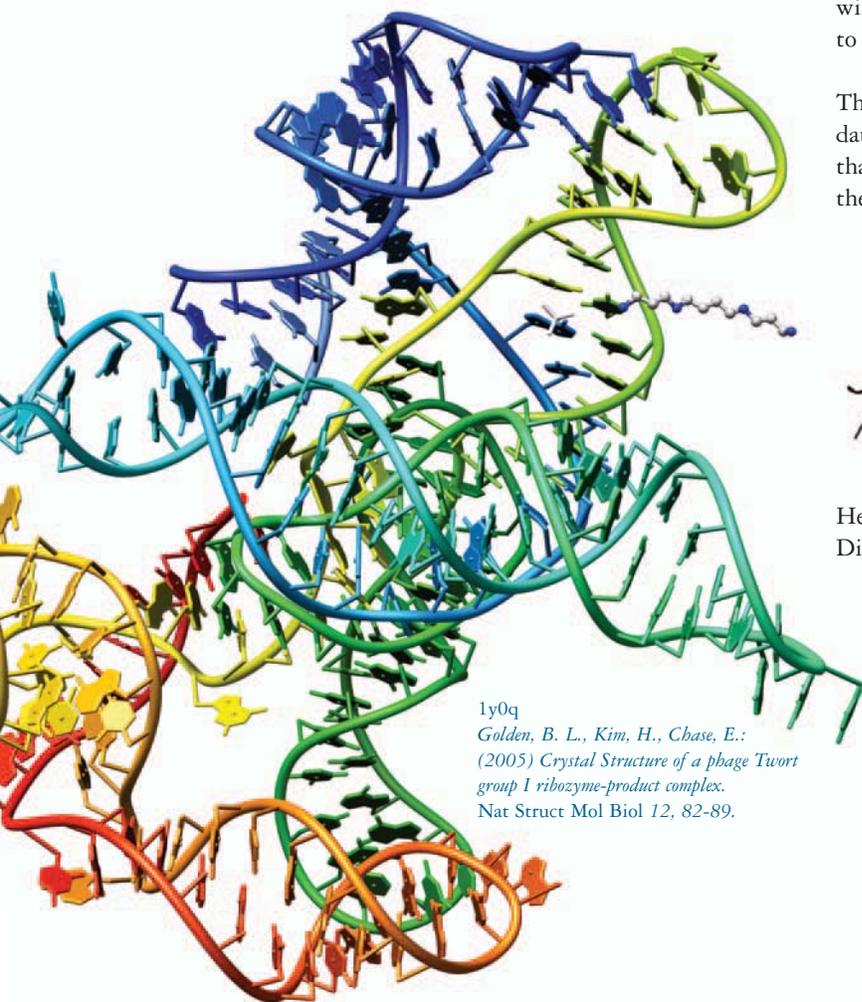
1y0q
Golden, B. L., Kim, H., Chase, E.:
(2005) Crystal Structure of a phage Twort
group I ribozyme-product complex.
Nat Struct Mol Biol *12, 82-89.*

What is the RCSB PDB?

The structures housed in the PDB range from small pieces of protein or DNA to complex machines, such as viruses and ribosomes. Each molecule plays a part in at least one biological process and each has value in helping scientists unravel the mysteries of life. Their structural shapes provide insight about fundamental biological processes and in some cases, about disease or drug interactions.

The PDB archives include a wide variety of medically important structures, including enzymes and other proteins associated with influenza, HIV, SARS and other viruses; parts of prion proteins (including the bovine form implicated in Mad Cow Disease); the amyloid peptide associated with Alzheimer's disease; and the p53 tumor-suppressor protein associated with a wide variety of human cancers.

The information in the PDB about these biological macromolecules holds significant promise for the pharmaceutical and biotechnology industries in the search for new drugs and in the effort to understand the mystery of human disease.

The data stored in the PDB include the three-dimensional coordinates for the structures of these biological macromolecules that have been determined experimentally using X-ray crystallography, NMR, and cryo-electron microscopy. Understanding what these structures look like and being able to 'see' how they bind to various ligands greatly aids in understanding how they function.

1u0n
Fukuda, K., Doggett, T., Laurenzi, I. J., Liddington, R. C., Diacovo, T. G. (2005) The snake venom protein botrocetin acts as a biological brace to promote dysfunctional platelet aggregation. Nat Struct Mol Biol *12, 152-159.*

The importance of the data related to three-dimensional structures was recognized early in the history of structural biology. The scientific community felt strongly that these data needed to be freely available to people in all fields of research. As a result, in 1971 the Protein Data Bank (PDB) was founded at Brookhaven National Laboratory[6] as the sole international repository for three-dimensional structure data of biological macromolecules. In 1999, management was assumed by the non-profit consortium the Research Collaboratory for Structural Bioinformatics[7].

This online repository contains the coordinates and related information for more than 31,000 structures.

SNAPSHOT – July 1, 2005
31535 released atomic coordinate entries

Molecule Type

28739	proteins, peptides, and viruses
1481	nucleic acids
1302	protein/nucleic acid complexes
13	carbohydrates

Experimental Technique

26932	diffraction and other
4603	NMR

The RCSB stores these data in a relational database that can be accessed using several different tools for searching and creating graphical reports. The RCSB PDB website also offers other resources for associated aspects of structural biology, including structural genomics, data representation formats, and downloadable software. The website includes a variety of materials for learning about structural biology and the PDB, and for interacting with the multifaceted interests of the PDB user communities.

Data residing in the PDB results from research projects all over the world. In recognition of the truly international nature of the PDB and the critical importance of maintaining that data in a single archive, the worldwide PDB (wwPDB) was established in 2003[8]. All wwPDB sites share responsibilities for data deposition, processing, and distribution of the PDB archives, and support the standardization of structure data (see also page 6).

PDB Users

As the sole repository for three-dimensional structure data of biological macromolecules, the PDB is a critical resource for academic, pharmaceutical, and biotechnology research. Structural biologists in industry and academia who focus on experimental structure determinations work with wwPDB staff to make these data available in the archives. Researchers from a wide variety of disciplines then use these data in their own studies, which can range from structural genomics to computational biology to structural biology and beyond.

Students and teachers also depend upon the RCSB PDB's collection of resources to help elucidate various subjects related to biological structure. Resources available include RCSB tools to find and visualize proteins, an education-centered webpage with materials for all levels of understanding, and the popular Molecule of the Month feature that regularly highlights different molecules from the PDB (page 13).
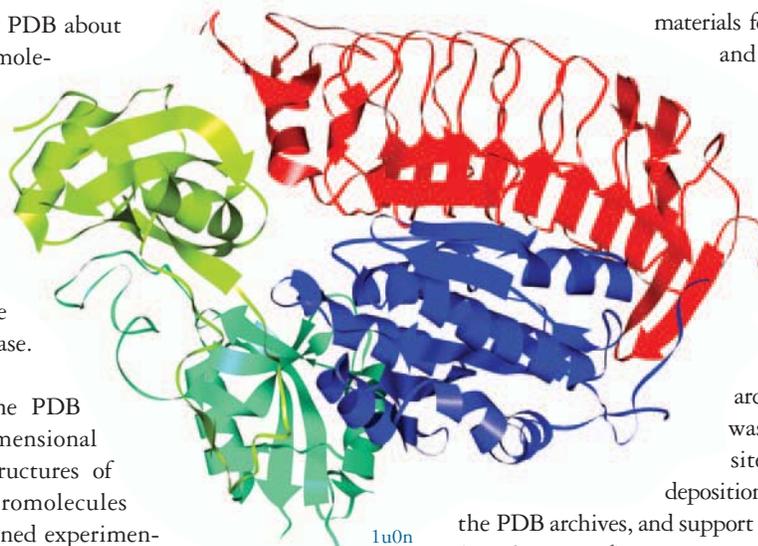
What is the RCSB PDB?

The RCSB PDB Team

The RCSB PDB Team Mission

The RCSB PDB seeks to enable science worldwide by offering a variety of resources to improve the understanding of structure-function relationships in biological systems. The RCSB PDB believes that the availability of consistent, well-annotated three-dimensional data will facilitate new scientific advances. For the data to be truly useful, the RCSB PDB must deliver it in a timely and efficient manner. To fulfill this mission, the capabilities of the RCSB PDB are continually being upgraded and significantly extended.

The RCSB also works with the wwPDB to maintain uniform archives of macromolecular structure data that are freely and publicly available to the global community.

RCSB PDB Management

The RCSB PDB is managed by Rutgers, The State University of New Jersey and the San Diego Supercomputer Center (SDSC), an organized research unit of the University of California, San Diego (UCSD) – two member institutions of the RCSB. The group from the Center for Advanced Research in Biotechnology/University of Maryland Biotechnology Institute/National Institute of Standards and Technology is no longer part of the RCSB consortium.

The Project Leaders manage the overall operations. Helen M. Berman, a Board of Governors Professor of Chemistry and Chemical Biology at Rutgers, is the Director of the RCSB PDB. She was part of the original team that developed the PDB at Brookhaven National Laboratory, and is the founder of the Nucleic Acid Database. Co-Directors John Westbrook, Research Associate Professor of Chemistry at Rutgers University and Philip E. Bourne, Professor of Pharmacology at UCSD oversee activities at their respective sites. Allison Clarke is the Operations Coordinator, and Judith L. Flippen-Anderson is the Outreach Coordinator.

RCSB PDB Leadership Team (from left to right): Allison Clarke, Dr. Philip E. Bourne, Dr. Helen M. Berman, Dr. John Westbrook, and Judith L. Flippen-Anderson

Funding

wwPDB: the worldwide PDB

The longstanding collaboration between the RCSB PDB and the Macromolecular Structure Database at the EMBL's European Bioinformatics Institute (MSD-EBI), and Protein Data Bank Japan (PDBj) was formalized in 2003. wwPDB underscores the international character of the PDB and ensures that the PDB archives will remain uniform. All three organizations serve as deposition, data processing and distribution sites of the PDB archives, with the RCSB maintaining the definitive FTP archive of files. Each wwPDB site provides its own view of the primary data, thus providing a variety of tools and resources for the global community. Additionally, wwPDB members collaborate on key projects essential to the maintenance of the PDB archives, such as data uniformity. Recent projects have included the distribution of yearly snapshots of the archives and the PDBML/XML representation of PDB data[9] (page 10).

The wwPDB receives guidance from an advisory board that it meets with on a yearly basis.

Advisory Committee

The RCSB PDB solicits the advice and guidance of its advisory committee. The PDB Advisory Committee (PDBAC) is chaired by Stephen K. Burley (Chief Scientific Officer and Senior Vice President, Research, Structural GenomiX). Its members are an international team of experts in X-ray crystallography, NMR, cryo-electron microscopy, bioinformatics, and education. They meet yearly with the RCSB PDB to review recent progress and to help plan for the ever-changing future.

The current membership includes Frank Allen (Executive Director, Cambridge Crystallographic Data Centre), Edward N. Baker (Professor, University of Auckland), Manju Bansal (Professor, Indian Institute of Science), Wah Chiu (Professor, Baylor College of Medicine), Paul Craig (Professor, Rochester Institute of Technology), Juli Feigon (Professor, University of California, Los Angeles), Andrzej Joachimiak (Director, Structural Biology Center at Argonne National Laboratory), Robert Kaptein (Professor, Utrecht University), Anthony J. Pawson (Professor, University of Toronto), Seth Pinsky (Director, Discovery Informatics, Abbott Laboratories), Andrej Sali (Professor, University of California, San Francisco), David Searls (Senior Vice-President of Bioinformatics, GlaxoSmithKline), and Cathy Wu (Professor, Georgetown University Medical Center).

Special thanks goes to PDBAC charter members who have recently ended their terms of service: Nobuhiro Go (Japan Atomic Energy Research Institute), Barry Honig (Professor, Columbia University), Sung-Hou Kim (Professor, University of California, Berkeley), and Judith Voet (Professor, Swarthmore College).

RCSB PDB Services: Data Deposition

Data Input: Deposition, Validation, and Annotation

A key component of the PDB is efficient capture (deposition) and curation (validation and annotation) of experimental structure data. Experiments using the techniques of X-ray crystallography, nuclear magnetic resonance (NMR), cryo-electron microscopy, and other methods produce the data that are deposited in the PDB. Scientists can contribute their data using tools available from wwPDB partners RCSB (US), PDBj (Japan) and MSD-EBI (United Kingdom). Data processed at the other wwPDB sites are forwarded to the RCSB for inclusion in the PDB archives.

The deposition tool ADIT (AutoDep Input Tool) is available online at RCSB and PDBj, and is also available as a software download for stand alone desktop use. ADIT provides access 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 (mmcif.pdb.org). mmCIF is an ontology of more than 3,500 terms defining macromolecular structure and related experiments.[10] (page 12)

1u77

Khademi, S., O'Connell III, J., Remis, J., Robles-Colmenares, Y., Miercke, L. J. W., Stroud, R. M. (2004) Mechanism of ammonia transport by Amt/MEP/Rh: structure of AmtB at 1.35 Å. Science 305, 1587-1594.

As soon as a structure is deposited using ADIT, it is assigned its own unique PDB ID. Staff then perform checks and annotate the deposited data. Validation reports and the completed file in mmCIF and PDB formats are sent to the depositor for review. Depositors also have the option of independently performing many of these checks using the validation option of ADIT. When finalized, the complete entry, including its status information and PDB ID, is loaded into the core relational database. RCSB PDB staff complete this entire process with an average turnaround of less than two weeks. Depending upon the hold status selected by the depositor, data release occurs when a depositor gives approval to the annotated entry, the hold date has expired, or the journal article has been published. As of May 2003, there is a one-year limit on the length of a hold period, including structures to be released upon publication of the related journal article. If the citation for a structure is not published within the one-year period, depositors are given the option to release the structure or withdraw the deposition.

Software for Deposition and Validation

The RCSB PDB has developed a variety of tools to facilitate data deposition and validation by the authors of structures, including pdb_extract, the Validation Suite and Server, Ligand Depot, and ADIT.

pdb_extract[12] helps depositors automatically prepare crystal structure depositions. This software tool extracts information about data collection, phasing, density modification, and the final structure refinement from the output files produced by many applications used for structure determination. The collected information is organized into a file that is ready for ADIT deposition. Fewer data items have to be manually entered by the depositor – saving time and minimizing errors. pdb_extract can be downloaded in source and binary versions, and is also part of the CCP4i interface (www.ccp4.ac.uk).

The Validation Suite and Server[13] can be used to check the coordinate format and validate the overall structure before deposition. The validation report contains geometrical and experimental checks from the programs MolProbity,[14] NUCheck,[15] PROCHECK,[16] and SFCHECK.[17] Sequence/coordinate alignment, missing and extra atoms or residues, and data inconsistencies are also reported.

Ligand Depot[18] is a data warehouse that integrates databases, services, tools and methods to provide chemical and structural information about the small molecules bound to macromolecules in the PDB. It can be used to find codes for existing ligands, link to entries with a particular ligand, and search for substructures.

Information missing from uploaded files can be added using the ADIT tool editor. mmCIF files written from the desktop version of ADIT can be uploaded into the web version for deposition.

All of these programs are updated regularly with downloads available for a variety of platforms.

RCSB PDB Services: Data Deposition

Statistics

During the period covered by this report, 6013 files were deposited to the wwPDB from around the world. Of these structures, approximately 81% were deposited with experimental data. Sequence data for about 57% of the depositions were released prior to the structure's release. More than half of the structures deposited to the PDB were processed by the RCSB PDB.

ADIT Depositions (RCSB PDB and PDBj; 5119 structures): July 2004 - June 2005
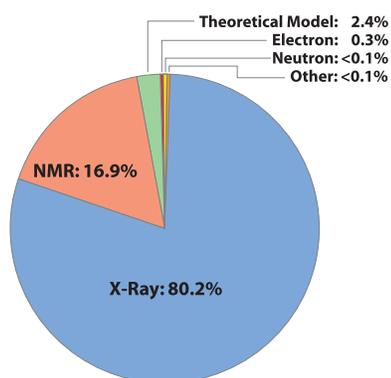
EXPERIMENTAL SOURCE	CONTINENT OF ORIGIN	RELEASE STATUS

EXPERIMENTAL SOURCE

Theoretical Model: 2.4%
Electron: 0.3%
Neutron: <0.1%
Other: <0.1%
NMR: 16.9%
X-Ray: 80.2%

CONTINENT OF ORIGIN

South America: 0.5%
Australia & New Zeland: 2%
Africa: <0.1%
Asia: 22%
Europe: 20.1%
North America: 55.4%

RELEASE STATUS

Release: 15.3%
HOLD: 14.9%
HPUB: 69.8%
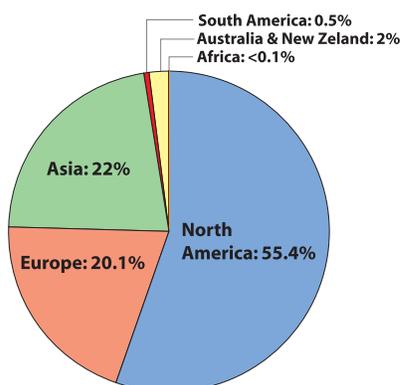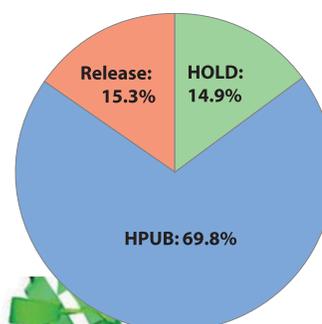
Annotators work to represent PDB data in the best possible way by ...

- Reviewing entry for self-consistency
- Matching given title to structure
- Correcting format errors in data and coordinates
- Checking sequence (using BLAST[11])
- Inserting sequence database reference
- Providing protein name and synonyms
- Checking scientific name
- Confirming ligand nomenclature
- Adding biological unit information
- Checking entry visually
- Generating validation reports
- Finding citation references with PubMed[11]
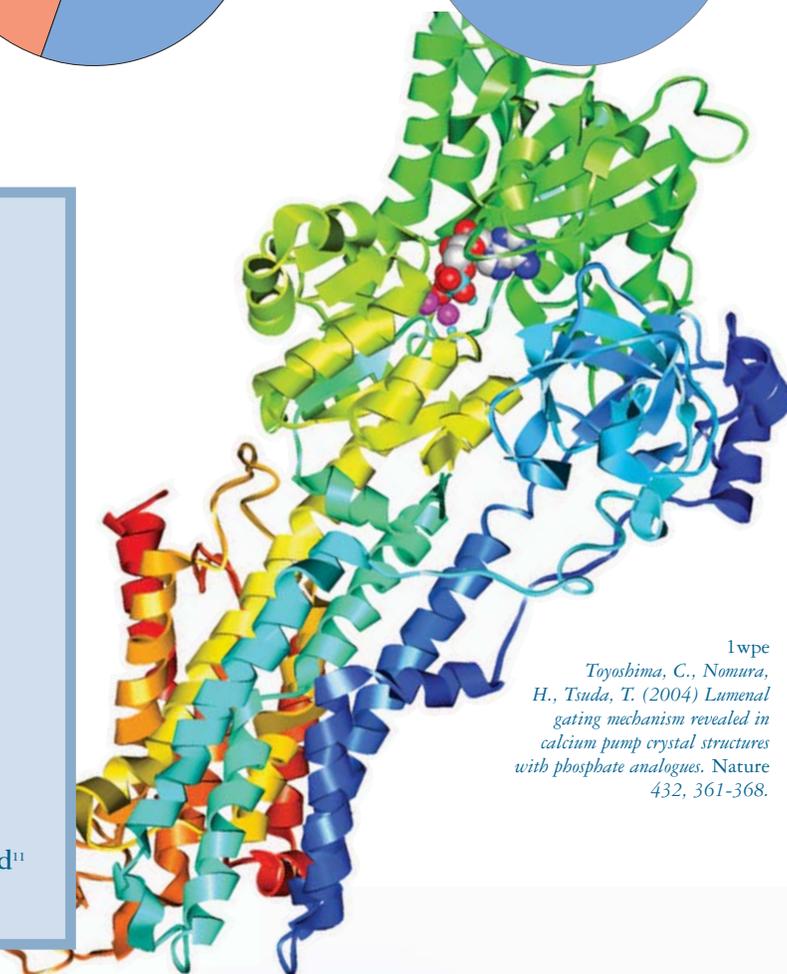
1wpe
Toyoshima, C., Nomura, H., Tsuda, T. (2004) Lumenal gating mechanism revealed in calcium pump crystal structures with phosphate analogues. Nature 432, 361-368.

RCSB PDB Services: Data Query, Reporting & Access

Data Distribution and Access

RCSB PDB services and data from the PDB archives are freely available through the internet. The main website at **www.pdb.org** receives approximately 270,000 hits per day from all over the world. On average, more than two files are downloaded each second. Additionally, there are five RCSB PDB mirror sites around the world at RCSB-Rutgers (US), Osaka University (Japan), the National University of Singapore (Singapore), the Cambridge Crystallographic Data Centre (United Kingdom), and the Max Delbrück Center for Molecular Medicine (Germany). The RCSB PDB website is maintained 24 hours a day, seven days a week.

New structures are added to the PDB holdings by 1:00 a.m. Pacific Time every Wednesday. The structures included in each release are highlighted on the homepage and clearly defined on the FTP site. During the period of this report, 5590 coordinate files were released into the archives, along with 3992 structure factor files, and 601 NMR constraint files.

Data Query and Reporting

Visitors to the RCSB PDB website can perform simple and complex queries on more than 31,000 structures and explore more than 800 curated webpages.

The goal of the RCSB PDB website is to make every structure accessible and comprehensible to all of our user communities – students and teachers, depositors, and all data users, including structural biologists, software developers, and others.

Whether searching for individual or multiple structures, users have a variety of options for searching and viewing structures. Each individual entry has a page that provides summary information, static and interactive images of the molecule, and links to other sources. A variety of reports can be created for a group of structures resulting from any query. Options to refine the query or create tabular reports from the search results are also available. A PDB or mmCIF format file for any structure can be downloaded as plain text or in one of several compressed formats from the PDB website. Files may also be downloaded from the PDB FTP server.

In July 2004, the RCSB PDB released a newly designed website with a reengineered database (**pdbbeta.rcsb.org**) for public beta testing alongside the current production site. Both sites will be updated regularly, until the beta site moves into production in 2006.

The new database design uses an industry-standard system that implements the PDB exchange data dictionary schema. The underlying database consists of mmCIF files produced by wwPDB member sites. Since the data being used is uniform across the archives and self-consistent, the number and type of queries across the archives are improved. The results of these queries are also more reliable and detailed.

The redesign of the website offers improved accessibility and navigation, and allows for the rapid addition of new features and resources to the RCSB PDB.

The new site provides improved ligand searching; a clear distinction between the reported primary and derived data; and the integration of even more data resources, such as SNP,[11] Pfam,[19] and Superfamily[20] assignments to genomes. Structural genomics has been given a specialized portal that is designed to provide improved access to data from these projects and from the TargetDB and PepcDB databases. An integrated help system is available to guide users through the RCSB PDB website, database, and structure data.

Thanks to the feedback sent from users, this website is constantly being refined and enhanced. Feedback comes from the PDB help desk, conference attendance, focus groups, advisory meetings, and other personal interactions between the users of the PDB and RCSB staff.

DAILY AVERAGE ACCESS STATISTICS FOR www.pdb.org
January 2000 – July 2005

RCSB PDB Services: Data Query, Reporting & Access

PDBML/XML Data Uniformity Files

After an extended period of beta testing, PDB data files are now available in XML format with each weekly update.

This XML format, called the PDB Markup Language (PDBML)[9], was developed as a wwPDB initiative. The PDBML schema description (pdbml.pdb.org) was produced by a direct translation of the PDB exchange data dictionary (mmcif.pdb.org). PDBML data files are produced by software translation of the data files in mmCIF format and are equivalent in content to these files. PDBML data files are provided in three forms: fully marked-up files, files without atom records, and files with a space efficient encoding of atom records.

Time-stamped Copies of PDB Archives

As part of a wwPDB initiative, time-stamped snapshots of the PDB archives will be added each year to ftp://snapshots.rcsb.org/. It is hoped that these snapshots will provide readily identifiable data sets for research on the PDB archives.

The first snapshot directory contains the exact and complete contents of the FTP archives as it appeared on January 6, 2005. These data, along with data in PDBML/XML format, were also made available to PDB users on eight DVDs. The yearly DVD set has replaced the CD-ROM distribution.

Browsing for Structures by Their Relationship to Disease, Molecular Function, Biochemical Process, or Cellular Location

A major feature of the new site is the ability to locate structures in the PDB by browsing through categories related to disease, molecular function, biochemical process, or cellular location. Organized in hierarchical "trees", the selection of one "branch" can lead to many other branches and structure types.

Browsers are available to navigate structures based upon the following classifications:
- **Gene Ontology** - biological process, cell component, and molecular function [21]
- **Enzyme Classification** - using EC nomenclature (www.chem.qmw.ac.uk/iubmb/enzyme)
- **Source Organism** - organisms in the NCBI Taxonomy database [11]
- **Disease** - disease categories and diseases in humans [11, 22]
- **Genome** - representatives of bacteria, archaea and eukaryota and several virus and organelle genomes [11]
- **SCOP** - comprehensive description of the evolutionary and functional relationships [23]
- **CATH** - a novel clustering of proteins at four major levels [24]

Users can explore each category's hierarchy, view the number of associated PDB structures, and search for specific related structures. For example, the Disease Browser can be used to look at diseases involving the nervous system, such as fragile X syndrome, Tay-Sachs, or Alzheimer Disease. Selecting a disease of interest will return all of the structures associated with that disease. The resulting list of structures can be sorted, downloaded, used to create a tabular report (*e.g.* citation or sequence information), or further refined by combining the search with another query.

Related Resources

Structural Genomics

Structural genomics is a worldwide initiative aimed at determining a large number of protein structures in a high throughput mode. During the period of this report, the first phase of the NIH-funded Protein Structure Initiative (PSI) reached the milestone of having solved over 1,000 structures, with the number of structures determined through genomics projects worldwide approaching 2,000. As the PDB is the repository for all of these structures, the RCSB PDB is actively involved in developing the informatics infrastructure needed for these projects, including the maintenance of data dictionaries describing these experiments. Deposition of structures generated by the genomics efforts has been facilitated by software that helps integrate data from structure determination packages and by close interactions between the depositors and annotation staff.

Efficient structure solution on a genomic scale requires a centralized coordination effort. Current and readily available information about the progress of protein production and structure solution is essential. The Target Registration Database (TargetDB, **target.pdb.org**)[25] collects and provides these data. It currently contains almost 89,000 sequences. Reports providing the experimental status of targets and sequence redundancy statistics are generated weekly. The Protein Expression Purification and Crystallization Database (PepcDB, **pepcdb.pdb.org**) extends the content of TargetDB with status history, stop conditions, reusable text protocols, and contact information. Both resources collect data from the NIH PSI projects and other structural genomics centers worldwide.

These database resources are used to prepare online summary reports about structural genomics projects worldwide, including target lists, target status progress, targets in the PDB, and sequence redundancy analyses. The structural genomics portal available from the RCSB PDB (**sg.pdb.org**) also provides a look at the distributions of different protein functions across the PDB archives using several schemas for classifying protein structure.

The RCSB PDB is also involved with the structural genomics initiatives through active participation in task forces, meetings, and individual interactions with each of the structural genomics centers.

Physical Archives

The PDB Physical Archives contain the paper documents, magnetic tapes, and other materials generated by the PDB since its beginnings at Brookhaven National Laboratory in 1971. The majority of the paper materials are stored in an environment-controlled archive facility in the Rutgers campus library system. The magnetic media have been transferred to a standard electronic media and stored in a local, searchable, database. The site now contains over 13 million files and consumes 690 Gbytes of storage.

The overall goal of the PDB Physical Archives is to preserve not only the original data submitted by the depositors, but also the records associated with all of 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, and aiding in uniformity and value-added annotation.

Recent Publications

Berman, H.M., and Westbrook, J. (2004) The impact of structural genomics on the Protein Data Bank. *Am J Pharmacogenomics* **4**, 247-252.

Chen, L., Oughtred, R., Berman, H.M., and Westbrook, J. (2004) TargetDB: A target registration database for structural genomics projects. *Bioinformatics* **20**, 2860-2862.

Deshpande, N., Addess, K.J., Bluhm, W.F., Merino-Ott, J.C., Townsend-Merino, W., Zhang, Q., Knezevich, C., Xie, L., Chen, L., Feng, Z., et al. (2005) The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema. *Nucleic Acids Res* **33**, D233-D237.

Dutta, S., and Berman, H.M. (2005) Large macromolecular complexes in the Protein Data Bank: A status report. *Structure* **13**, 381-388.

Dutta, S., Burkhardt, K., Bluhm, W.F., and Berman, H.M. (2005) Using the tools and resources of the RCSB Protein Data Bank. *Current Protocols in Bioinformatics*, 1.9.1-1.9.40.

Feng, Z., Chen, L., Maddula, H., Akcan, O., Oughtred, R., Berman, H.M., and Westbrook, J. (2004) Ligand Depot: a data warehouse for ligands bound to macromolecules. *Bioinformatics* **20**, 2153-2155.

Pajon, A., Ionides, J., Diprose, J., Fillon, J., Fogh, R., Ashton, A.W., Berman, H., Boucher, W., Cygler, M., Deleury, E., Esnouf, R., Janin, J., Kim, R., Krimm, I., Lawson, C. L., Oeuillet, E., Poupon, A., Raymond, S., Stevens, T., van Tilbeurgh, H., Westbrook, J., Wood, P., Ulrich, E., Vranken, W., Xueli, L., Laue, E., Stuart, D. I. Henrick, K. (2005) Design of a data model for developing laboratory information management and analysis systems for protein production. *Proteins* **58**, 278-284.

Westbrook, J., Ito, N., Nakamura, H., Henrick, K., and Berman, H.M. (2005) PDBML: The representation of archival macromolecular structure data in XML. *Bioinformatics* **21**, 988-992.

Yang, H., Guranovic, V., Dutta, S., Feng, Z., Berman, H.M., and Westbrook, J. (2004) Automated and accurate deposition of structures solved by X-ray diffraction to the Protein Data Bank. *Acta Crystallogr D Biol Crystallogr* **60**, 1833-1839.

Related Resources

Data Dictionaries: mmCIF and the PDB Exchange Dictionary

The RCSB PDB uses macromolecular Crystallographic Information File (mmCIF) data dictionaries to describe the information content of PDB entries. The mmCIF dictionary is an ontology of more than 3,500 terms defining macromolecular structure and its related experiments.

The PDB exchange dictionary consolidates content from a variety of dictionaries and includes extensions to describe NMR, cryo-electron microscopy, and protein production data. PDB data processing, data exchange, annotation, and database management operations all make heavy use of the mmCIF data format and the content of the PDB exchange dictionary.

PDB entries can be downloaded from the RCSB PDB website or by ftp in mmCIF, PDB, and PDBML/XML formats. Software tools are available for preparing and editing files for new depositions, and for converting mmCIF data files to PDB and PDBML/XML formats.

NMR Structures

As approximately 15% of the structures in the PDB archives have been determined by NMR experiments, the RCSB PDB works to provide relevant information about these structures.

The RCSB PDB collaborates closely with the BioMagResBank (BMRB; **www.bmrb.wisc.edu**)[26], the repository for NMR experimental data.

A version of ADIT that collects NMR experimental data – ADIT-NMR – is now in use at the BMRB. ADIT-NMR will be expanded to include coordinate data. It uses an mmCIF-like dictionary (NMRIF) that was derived from the NMR STAR deposition form used by the BMRB. The ADIT system has been extended to display deposition information that closely models the current BMRB NMR STAR data presentation, and is now being used by BMRB to collect NMR experimental data.

The RCSB PDB website can be used to create searches and reports centered on data specific to NMR structures. These featured items include NMR experiment type, refinement method, selection criteria, spectrometer manufacturer, spectrometer model, and sample conditions.

RCSB PDB staff members attend meetings and work closely with the NMR community. A joint RCSB PDB/BMRB NMR Task Force is being formed as an advisory group for issues relating to NMR deposition.

Community Interactions

With a diverse user community of depositors, data users, students, teachers, and the general public, the RCSB PDB relies on a variety of methods to keep in contact with all members of the community. The goals of our outreach efforts are to provide information about the resource, to gain feedback for further development of the RCSB PDB, and to provide materials that promote a broader understanding and scientific literacy of structural biology. We place special emphasis on reaching out to new users, students, educators, and the general public.

RCSB PDB advancements are highlighted online, in print, and in-person. Activities and new features are announced regularly on the RCSB PDB homepage. We distribute a quarterly newsletter as well as a variety of online and print materials about RCSB PDB resources, including tutorials and educational materials. Various projects and developments are also described in books and journals (page 11).

Communication with our users is promoted through accessibility – we maintain an active help desk and have a strong presence at meetings through presentations, workshops, and exhibit booths. Staff rapidly responds to general and specific inquiries sent to our help desks. In addition, the **pdb-l@rcsb.org** listserv offers a forum for discussion among members of the PDB community.

The RCSB PDB has participated in several meetings **(images 1 & 2)** to interact with depositors (*e.g.* American Crystallographic Association (ACA), Frontiers of NMR in Molecular Biology), data users (Intelligent Systems for Molecular Biology), the
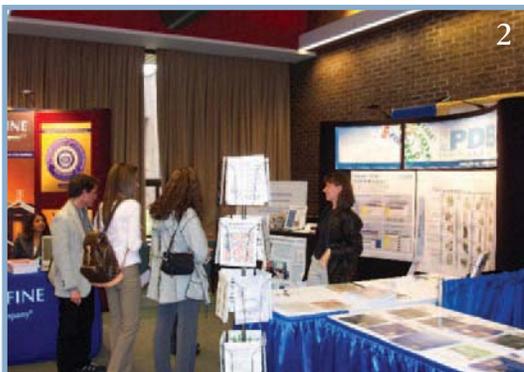
Related Resources

structural genomics community (International Conference on Structural Genomics), and educators (National Science Teachers Association).

Our Art of Science traveling exhibit looks at the beauty inherent in protein structures. It displays images of molecules in the PDB, including the pictures available from Structure Explorer pages and from Molecule of the Month features. Since its inception in a gallery space at Rutgers University, the show has traveled to many places, including Texas A&M University (image 3); EMBL-Hamburg, Germany (image 4); University of Wisconsin-Madison; California State University, Fullerton; Purdue University; and Hyderabad, India.

The 2004 RCSB PDB Poster Prize recognized student poster presentations involving macromolecular crystallography at each of the meetings of the International Union of Crystallography's Regional Associates–the ACA, the Asian Crystallographic Association, and the European Crystallographic Association. The prize was also awarded at the Annual International Conference on Research in Computational Molecular Biology (RECOMB).

Molecule of the Month Series

Highlighted on the RCSB PDB homepage, the Molecule of the Month series presents short accounts of selected molecules from the PDB. Each installment includes an introduction to the structure and function of the molecule and relates the molecule to human health and welfare. Suggestions for viewing structures on the RCSB PDB website and for additional reading are provided. Produced and illustrated by David S. Goodsell since 2000, the Molecule of the Month is a proven resource for the classroom.

Other Educational Resources

The RCSB PDB provides vital resources for teachers, students, and anyone with a general interest in structural biology. (image 5) Tools for finding and visualizing proteins are used in a variety of ways, including downloading molecular images, exploring the links to information found in journals, and trying different keyword queries to locate specific proteins.

Curricular materials used by educators are solicited and posted on the RCSB PDB website. The Education pages provide resources for learning about proteins and nucleic acids, protein documentaries, and suggested reading materials and links. Tutorials and help documentation are also available, with enhanced user help features available on the RCSB PDB beta test site.

The RCSB PDB Newsletter regularly features interviews with members of the community and descriptions of how the PDB is used in all levels of education. Recent Education Corners in the newsletter have included a look at using visualization tools in the classroom and a protein modeling event at the Science Olympiad. (image 6)
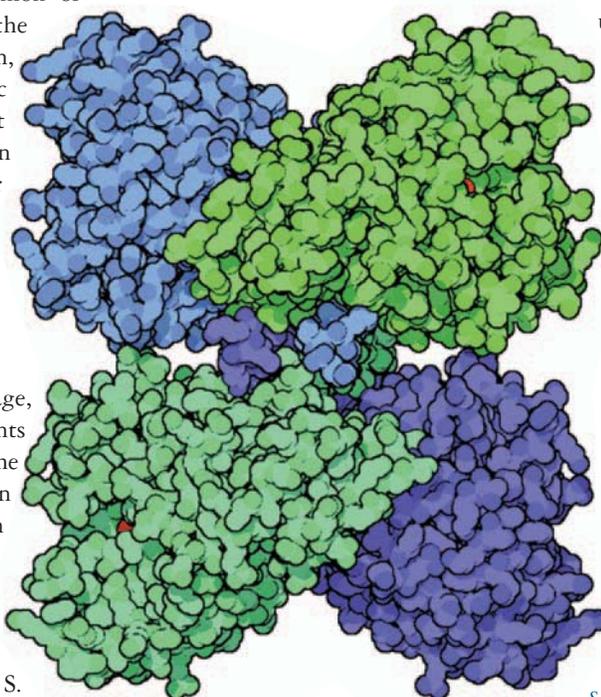
Phenylalanine Hydroxylase from the January 2005 Molecule of the Month feature:

2pah
Fusetti, F., Erlandsen, H., Flatmark, T., Stevens, R. C.: Structure of Tetrameric Human Phenylalanine Hydroxylase and its Implications for Phenylketonuria J.Biol.Chem. *273 pp. 16962-16967 (1998)*

1phz
Kobe, B., Jennings, I. G., House, C. M., Michell, B. J., Goodwill, K. E., Santarsiero, B. D., Stevens, R. C., Cotton, R. G., Kemp, B. E.: Structural basis of autoregulation of phenylalanine hydroxylase. Nat Struct Biol 6 *pp. 442-448 (1999)*

Related Resources

Workshops

The RCSB PDB has organized and sponsored a variety of workshops and symposia related to structural biology. The *Database Challenges in Biology* symposium (September 10, 2004) brought together experts in biological data management to describe data organization methods that enable scientists to derive new knowledge about structure and function.

Together with CCP4[27], the RCSB PDB held a workshop entitled *A Protein Crystallographic Toolbox: CCP4 Software Suite and RCSB PDB Deposition Tools* during the 2004 ACA meeting. The goal for this workshop was to teach depositors how to use some of the protein crystallography programs distributed with the CCP4 package and how to get the best from the PDB when they deposit their structures.

• Cryo-Electron Microscopy Workshop

Due to their complexity, structures determined by cryo-electron microscopy provide unique challenges to the RCSB PDB – these structures have many independent chains, complex symmetries, high variability in quality of atomic level description, and complicated routes to generating structures involving multiple types of experimental data. A recent workshop brought together electron microscopists, programmers and database experts to develop a community consensus on the data items needed for deposition of 3D density maps and atomic models derived from cryo-electron microscopy studies. The *Cryo-Electron Microscopy Structure Deposition Workshop* held at RCSB-Rutgers (October 23-24, 2004) examined the data items currently collected by EMDep (for the EM Data Bank at MSD-EBI) and ADIT for these depositions, and discussed desirable additions for the near future. The results included a revised data dictionary that is currently under final review, and recommendations for improvements in the areas of visualization, data mining and data integration. The workshop was sponsored by the RCSB PDB and the Computational Center for Biomolecular Complexes (C²BC; ncmi.bcm.tmc.edu/ccbc).
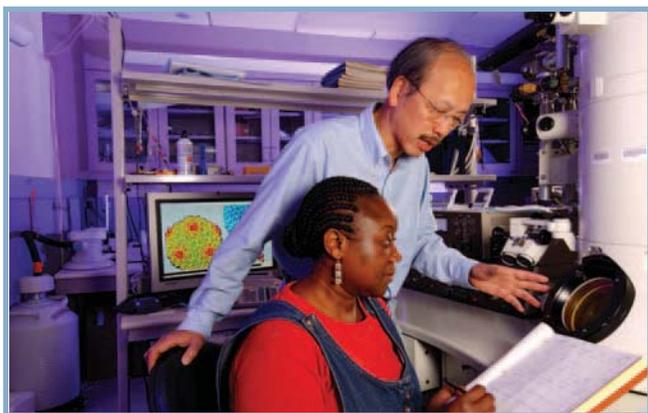
Joanita Jakana and Wah Chiu are with a JEM3000SFF electron cryomicroscope, a 300 kV electron microscope that can keep specimens at liquid helium temperature for cryo-electron microscopy experiments.

• Educational Workshops

The educational resources available from the RCSB PDB were highlighted during workshops aimed at educators. At the *X-rays, Molecules, and You* workshop held at the 2004 ACA meeting, prominent crystallographers and structural biologists introduced high school students and teachers from the Midwest to basic concepts in X-ray crystallography and structural databases to promote interest in structural biology, chemistry, and general science. It provided hands-on demonstrations for using the RCSB PDB database and growing protein crystals. The symposium *PDB in the Classroom and Beyond* was held as part of the "Education in the Biomolecular Sciences: The Next Generation" symposium at the American Society for Biochemistry and Molecular Biology's (ASBMB) 2005 Annual Meeting. Presentations discussed how the RCSB PDB is used as a teaching tool for a variety of subjects.

References

Molecular images were created using Chimera,[28] unless otherwise noted.

[1] Cornilescu, G., Cornilescu, C.C., Zhao, Q., Frederick, R.O., Peterson, F.C., Thao, S. and Markley, J.L. (2004) Solution structure of a homodimeric hypothetical protein, At5g22580, a structural genomics target from *Arabidopsis thaliana*. *J. Biomol. NMR*, **29**, 387-390.

[2] Deng, L., Starostina, N., Liu, Z.-J., Rose, J.P., Terns, R.M., Terns, M.P. and Wang, B.-C. (2004) Structure determination of fibrillarin from the hyperthermophilic archaeon *Pyrococcus furiosus. Biochem. Biophys. Res. Comm.*, **315**, 726-732.

[3] Schwarzenbacher, R., Jaroszewski, L., von Delft, F., Abdubek, P., Ambing, E., Biorac, T., Brinen, L.S., Canaves, J.M., Cambell, J., Chiu, H.J. et al. (2004) Crystal structure of a phosphoribosylaminoimidazole mutase PurE (TM0446) from *Thermotoga maritima* at 1.77 Å resolution. *Proteins*, **55**, 474-478.

[4] Sharma, V., Arockiasamy, A., Ronning, D.R., Savva, C.G., Holzenburg, A., Braunstein, M., Jacobs, W.R., Jr. and Sacchettini, J.C. (2003) Crystal structure of *Mycobacterium tuberculosis* SecA, a preprotein translocating ATPase. *Proc. Natl. Acad. Sci. USA*, **100**, 2243-2248.

[5] Chen, S., Jancrick, J., Yokota, H., Kim, R. and Kim, S.H. (2004) Crystal structure of a protein associated with cell division from *Mycoplasma pneumoniae* (GI: 13508053): a novel fold with a conserved sequence motif. *Proteins*, **55**, 785-791.

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

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

[8] Berman, H.M., Henrick, K. and Nakamura, H. (2003) Announcing the worldwide Protein Data Bank. *Nat Struct Biol*, **10**, 980.

[9] Westbrook, J., Ito, N., Nakamura, H., Henrick, K. and Berman, H.M. (2005) PDBML: The representation of archival macromolecular structure data in XML. *Bioinformatics*, **21**, 988-992.

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

[11] Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S., Helmberg, W. et al. (2005) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **33**, D39-45.

[12] Yang, H., Guranovic, V., Dutta, S., Feng, Z., Berman, H.M. and Westbrook, J. (2004) Automated and accurate deposition of structures solved by X-ray diffraction to the Protein Data Bank. *Acta Crystallogr. D Biol. Crystallogr.*, **60**, 1833-1839.

[13] Westbrook, J., Feng, Z., Burkhardt, K. and Berman, H.M. (2003) Validation of protein structures for the Protein Data Bank. *Meth Enz.*, **374**, 370-385.

[14] Davis, I.W., Murray, L.W., Richardson, J.S. and Richardson, D.C. (2004) MOLPROBITY: structure validation and all-atom contact analysis for nucleic acids and their complexes. *Nucleic Acids Res.*, **32**, W615-619.

[15] Feng, Z., Westbrook, J. and Berman, H.M. (1998) NUCheck. Rutgers University, New Brunswick, NJ.

[16] Laskowski, R.A., MacArthur, M.W., Moss, D.S. and Thornton, J.M. (1993) PROCHECK - a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.*, **26**, 283-291.

[17] Vaguine, A.A., Richelle, J. and Wodak, S.J. (1999) SFCHECK: a unified set of procedures for evaluating the quality of macromolecular structure-factor data and their agreement with the atomic model. *Acta Crystallogr. D Biol. Crystallogr.*, **55**, 191-205.

[18] Feng, Z., Chen, L., Maddula, H., Akcan, O., Oughtred, R., Berman, H.M. and Westbrook, J. (2004) Ligand Depot: A data warehouse for ligands bound to macromolecules. *Bioinformatics*, **20**, 2153-2155.

[19] Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S.R., Griffiths-Jones, S., Howe, K.L., Marshall, M. and Sonnhammer, E.L. (2002) The Pfam protein families database. *Nucleic Acids Res.*, **30**, 276-280.

[20] Madera, M., Vogel, C., Kummerfeld, S.K., Chothia, C. and Gough, J. (2004) The SUPERFAMILY database in 2004: additions and improvements. *Nucleic Acids Res.*, **32**, D235-239.

[21] The Gene Ontology Consortium. (2000) Gene Ontology: tool for the unification of biology. *Nature Genetics*, **25**, 25-29.

[22] NCBI. (1998-) Genes and disease [electronic resource]. National Center for Biotechnology Information, Bethesda (MD). Updated weekly. National Library of Medicine. Available from http://www.ncbi.nlm.nih.gov/books/bv.fcgi?call=bv.View..ShowSection&rid=gnd.preface.91.

[23] Conte, L., Bart, A., Hubbard, T., Brenner, S., Murzin, A. and Chothia, C. (2000) SCOP: a structural classification of proteins database. *Nucleic Acids Res.*, **28**, 257-259.

[24] Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B. and Thornton, J.M. (1997) CATH–a hierarchic classification of protein domain structures. *Structure*, **5**, 1093-1108.

[25] Chen, L., Oughtred, R., Berman, H.M. and Westbrook, J. (2004) TargetDB: A target registration database for structural genomics projects. *Bioinformatics*, **20**, 2860-2862.

[26] Doreleijers, J.F., Mading, S., Maziuk, D., Sojourner, K., Yin, L., Zhu, J., Markley, J.L. and Ulrich, E.L. (2003) BioMagResBank database with sets of experimental NMR constraints corresponding to the structures of over 1400 biomolecules deposited in the Protein Data Bank. *J. Biomol. NMR*, **26**, 139-146.

[27] CCP4. (1994) The Collaborative Computational Project Number 4 suite: programs for protein crystallography. *Acta Crystallogr. D Biol. Crystallogr.*, **50**, 760-763.

[28] Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C. and Ferrin, T.E. (2004) UCSF Chimera--a visualization system for exploratory research and analysis. *J. Comput. Chem.*, **25**, 1605-1612.

wwPDB Members

RCSB PDB: www.pdb.org

Macromolecular Structure Database at the European Bioinformatics Institute:
www.ebi.ac.uk/msd

Protein Data Bank Japan:
www.pdbj.org

RCSB PDB Partners

Rutgers, The State University of New Jersey, Department of Chemistry and Chemical Biology
610 Taylor Road
Piscataway, NJ 08854-8087

San Diego Supercomputer Center at the University of California, San Diego
9500 Gilman Drive
La Jolla, CA 92093-0537

RCSB PDB Leadership Team

Dr. Helen M. Berman, Director
Rutgers, The State University of New Jersey
berman@rcsb.rutgers.edu

Dr. Philip E. Bourne, Co-Director
San Diego Supercomputer Center
University of California, San Diego
bourne@sdsc.edu

Allison Clarke, Operations Coordinator
Rutgers, The State University of New Jersey
aclarke@rcsb.rutgers.edu

Judith L. Flippen-Anderson, Outreach Coordinator
Rutgers, The State University of New Jersey
flippen@rcsb.rutgers.edu

Dr. John Westbrook, Co-Director
Rutgers, The State University of New Jersey
jwest@rcsb.rutgers.edu

A list of current RCSB PDB Team Members is available at
www.rcsb.org/pdb/rcsb-group.html

1yce
Meier, T., Polzer, P., Diederichs, K., Welte, W., Dimroth, P. (2005)
Structure of the rotor ring of F-Type Na+-ATPase from
Ilyobacter tartaricus. Science *308,*
659-662.
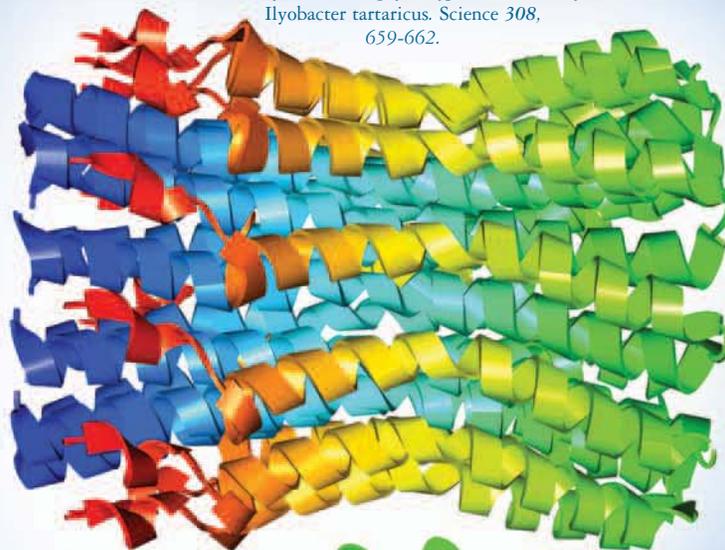
2bl2
Murata, T., Yamato, I.,
Kakinuma, Y., Leslie, A. G.
W., Walker, J. E. (2005)
Structure of the rotor of the
Vacuolar-Type Na-ATPase from
Enterococcus hirae. Science
308, 654-659.
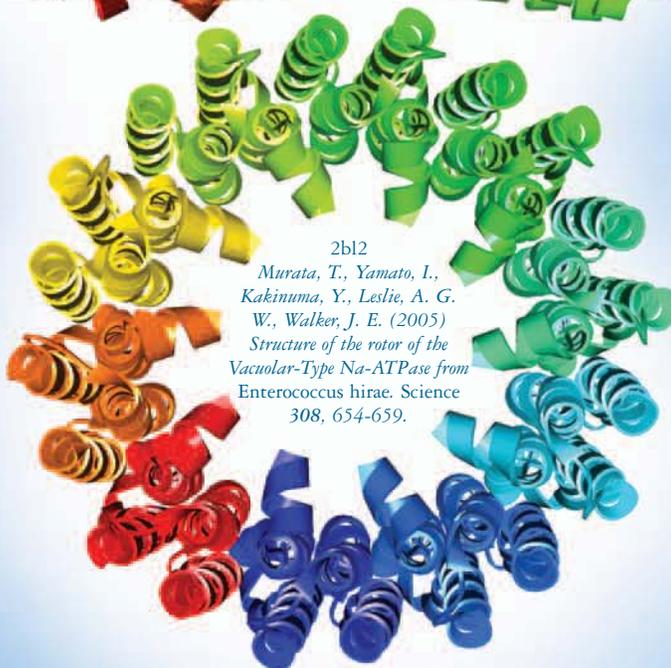
RCSB PDB

www.pdb.org
Send questions or comments to:
info@rcsb.org