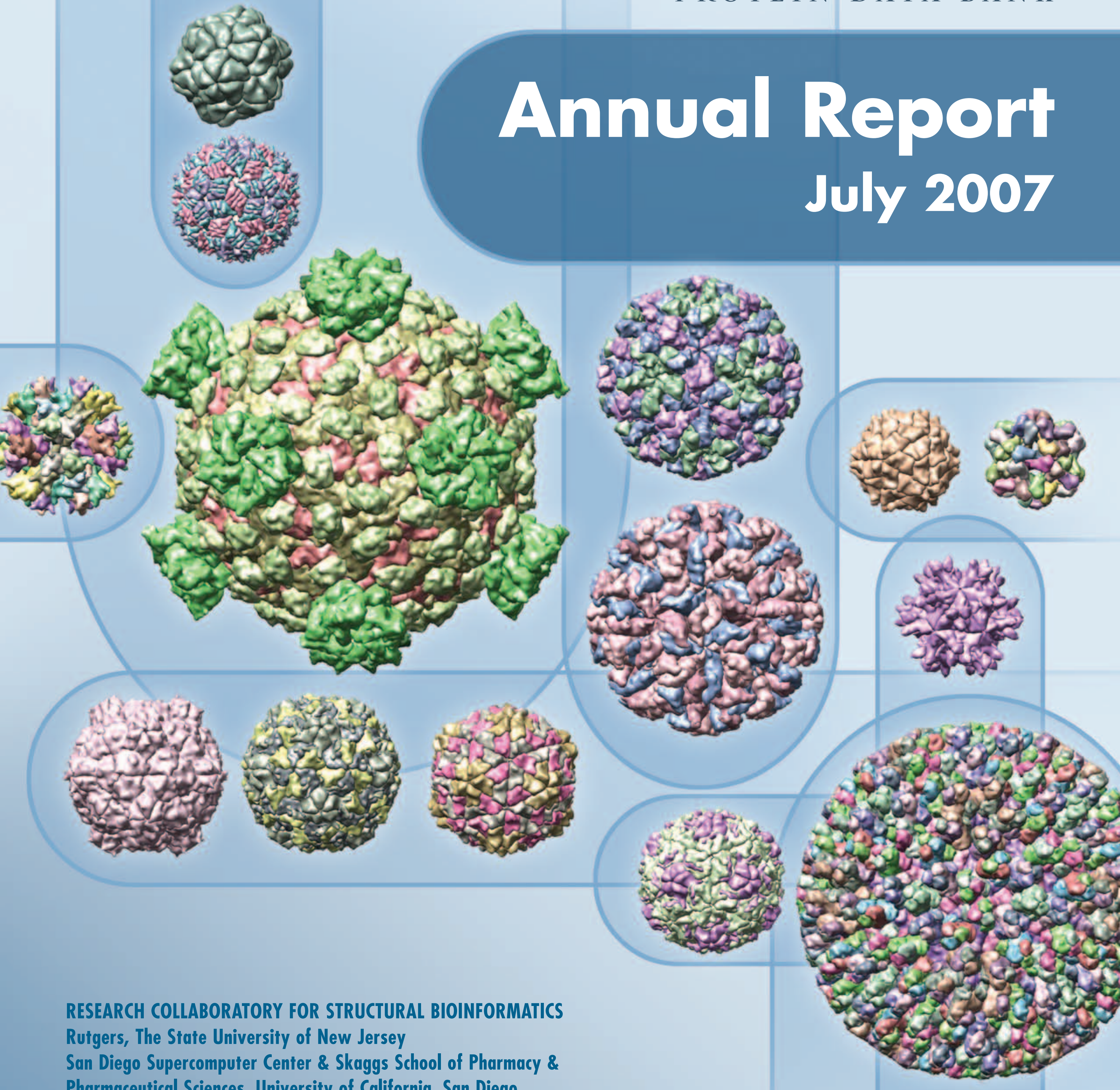
RCSB PDB
PROTEIN DATA BANK

Annual Report
July 2007



RESEARCH COLLABORATORY FOR STRUCTURAL BIOINFORMATICS
Rutgers, The State University of New Jersey
San Diego Supercomputer Center & Skaggs School of Pharmacy &
Pharmaceutical Sciences, University of California, San Diego

About the Cover

RCSB PDB PROTEIN DATA BANK

Annual Report
July 2007

RESEARCH COLLABORATORY FOR STRUCTURAL BIOINFORMATICS
Rutgers, The State University of New Jersey
San Diego Supercomputer Center & Skaggs School of Pharmacy & Pharmaceutical Sciences, University of California, San Diego

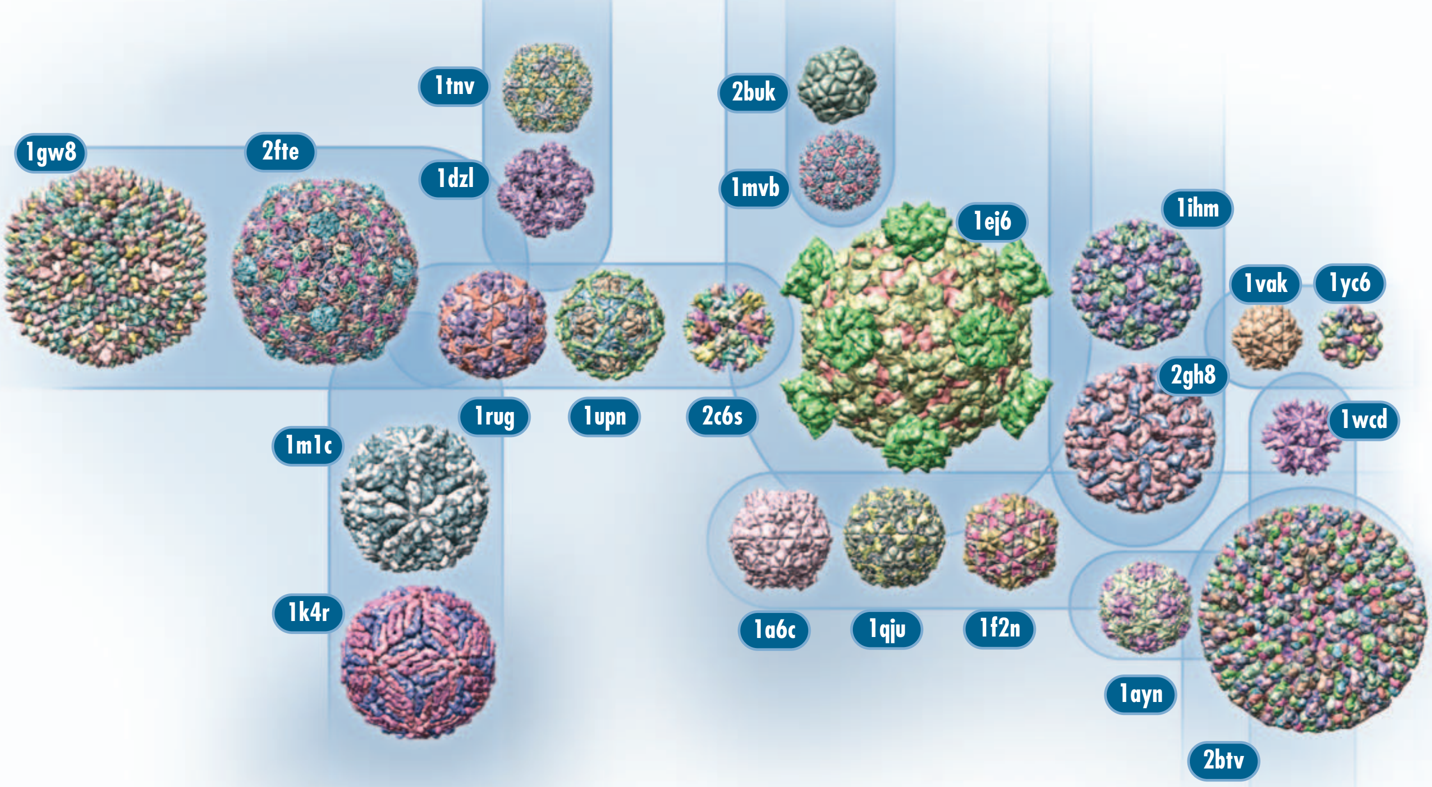
Viruses are tiny, simple machines, minimally composed of genetic material encased in a protein shell.

These virus shells, or capsids, that surround the viral genome are often highly symmetrical. Typically, only a portion of the complete virus structure, along with information about how to build the entire shell using symmetry operations, is deposited to the archive. However, the annotation of virus entries has not been consistent over the past two decades. In many cases, the symmetry information was missing, incorrect, or provided in a way that was difficult to interpret. Users would have to try to figure out how to build the full molecule on their own, with mixed results.

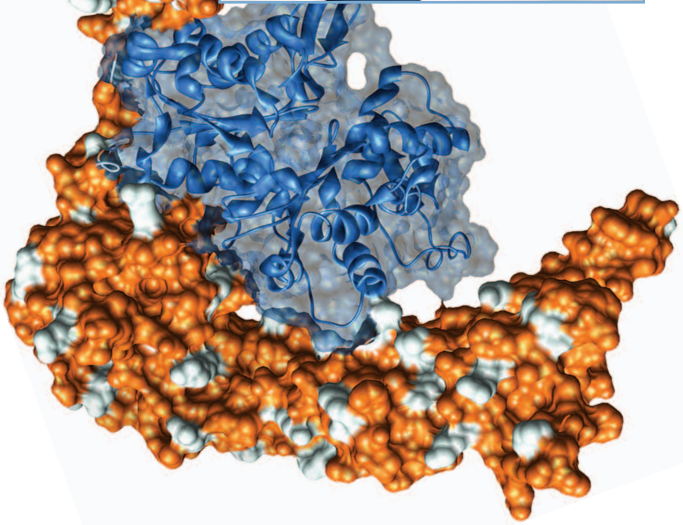
As part of the wwPDB Remediation Project, every virus entry in the PDB has been examined and updated. The complete virus can now be easily and consistently constructed based on the information included in the revised files.

The images and information about the viral capsids shown on the cover are now easily accessible from the RCSB PDB website.

Contents



Message from the Director



1y64: *T. Otomo, D.R. Tomchick, C. Otomo, S.C. Panchal, M. Machius, M.K. Rosen (2005) Structural basis of actin filament nucleation and processive capping by a formin homology 2 domain* Nature *433:488-494.*

Established in 1971, the Protein Data Bank (PDB) has been the sole archive for the 3D structures of biological macromolecules. At its beginning, the archive held information describing 7 crystal structures, with new depositions added when data were sent via punch card or tape through the mail.

More than 44,000 structures later, the PDB archive is available online as an important resource for chemistry, structural biology, computational biology, pharmacology, and education. Since its beginnings, the PDB has seen new experimental techniques used in structure determinations, such as NMR and electron microscopy, while X-ray crystallographic techniques continually evolve. These developments in the types of data contained in the PDB, along with inconsistently reported data, nomenclature, and other annotations, have limited the types of queries possible across the archive.

Over the past several years, members of the international organization that annotates and releases PDB data–the Worldwide Protein Data Bank–have worked together to review and remediate the archive. Virtually every structure in the archive has been enhanced or updated as part of this effort. In the summer of 2007, the PDB archive was updated to incorporate these remediated data. Now that the PDB archive has been made uniform, the archive has been opened up for further understanding and study.

The RCSB PDB website offers unique tools to mine the data contained in this archive. To utilize the benefits of the remediated data, the website has been updated to provide improved database searching and reporting capabilities, enhanced sequence information, and advanced access to ligand information.

These developments are highlighted in this annual report, along with descriptions of other projects that ensure that the RCSB PDB provides an ever-evolving portal for studying the structures of biological macromolecules and their relationships to sequence, function, and disease.

Helen Berman

Helen M. Berman
Director, RCSB PDB
Board of Governors Professor of Chemistry and Chemical Biology
Rutgers, The State University of New Jersey

About the RCSB PDB

The RCSB Protein Data Bank, administered by the nonprofit Research Collaboratory for Structural Bioinformatics (RCSB), aims to support worldwide scientific research by providing an essential library of protein and nucleic acid structures.

In addition to developing tools and systems for the deposition, annotation, and release of data in the PDB archive, the RCSB PDB supports a website where visitors can perform simple and complex queries on the data, analyze, and visualize the results. The RCSB PDB is used by researchers from a wide variety of disciplines, students, teachers, and the general public.

SNAPSHOT — JULY 1, 2007

44,320 released atomic coordinate entries

Molecule Type

40,730	proteins, peptides, and viruses
1,760	nucleic acids
1,795	protein/nucleic acid complexes
35	other

Experimental Technique

37,716	X-ray
6,367	NMR
149	electron microscopy
88	other

27,026	structure factor files
3,512	NMR restraint files

The Proteins and the Promise

Proteins are one of the main building blocks for living organisms. Proteins come in a variety of shapes, and the form of a protein corresponds to its function. For example, a protein that transports molecules across a membrane may take the form of a hollow tunnel with specific characteristics that allow it to be embedded in the membrane. The structures in the PDB range from tiny proteins and bits of DNA to complex molecular machinery like the ribosome, the cellular factory that assembles other proteins.

Knowledge of a protein's structure may help scientists deduce its role in human health or disease. In addition, this information is useful in drug development. For instance, knowledge of the shape of two HIV proteins–reverse transcriptase and HIV protease–allowed scientists to design drugs that prevent the HIV virus from multiplying. These drugs now form a part of life-prolonging therapy for HIV patients.

Among the over 44,000 proteins currently represented in the PDB are:

- Insulin, the protein deficient in diabetic patients
- p53 tumor suppressor, a protein often implicated in cancer
- Influenza proteins, structures which may help scientists design medicines to combat the flu
- Anthrax toxin, the disease-causing protein made by anthrax
- Prion proteins, misshapen proteins that are the cause of many diseases, including mad cow disease
- Amyloid peptide, a protein implicated in Alzheimer's disease

Solving the puzzle of a protein's shape requires advanced techniques and careful analysis. Scientists use methods such as X-ray crystallography, nuclear magnetic resonance (NMR), and 3D electron microscopy (3D EM) to acquire structural data.

When a scientist determines a new structure of a protein, this information is deposited in the PDB archive where it will be freely available. With over 5 million files downloaded from the FTP archive each month, this valuable data is shared around the globe, contributing to new studies and perhaps new benefits to human health.

36 Years of Protein Structures

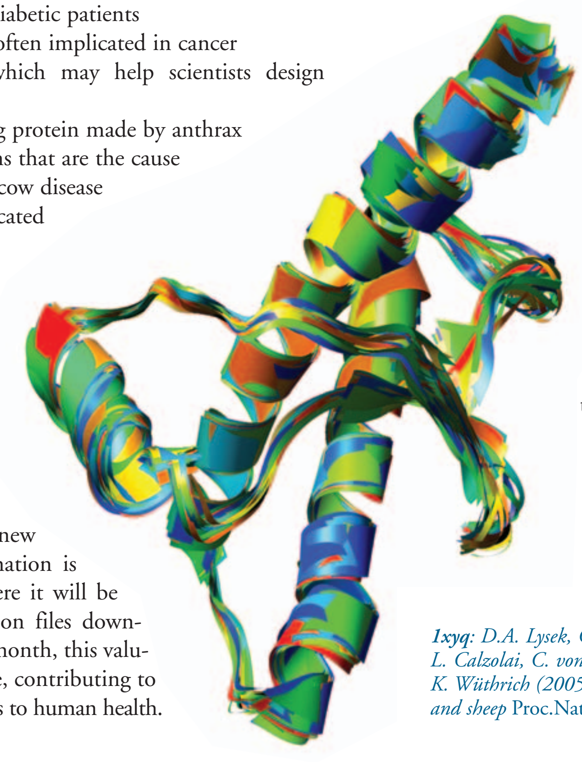
Almost as soon as it became possible to deduce the form of a protein, scientists realized that this important information must be shared. The PDB archive is the sole international repository for 3D structures of biological macromolecules. The resource includes proteins, nucleic acids (including DNA and RNA), and protein-nucleic acid complexes.

Today, the PDB archive receives almost 20 new structures *daily.* Curators then check each deposit for errors or omissions, ensuring a consistent and accurate entry for each protein.

Several resources make these structural data accessible. Visualization software allows users to "see" a 3D picture of a protein structure. Other tools help PDB users search databases, tabulate data, and analyze the results.

Founded at Brookhaven National Laboratory in 1971[1,2] as a library of data for all protein structures, the PDB archive became the responsibility of the RCSB PDB in 1999.[3] In 2003, the Worldwide PDB (wwPDB) was founded to recognize the international nature of the PDB archive.[4] wwPDB members share responsibility for annotating the data deposited, while providing different views to the data. As "archive keeper" for the wwPDB, the RCSB PDB controls the central repository of the PDB archive.

The mission of the RCSB PDB is to foster scientific advance by providing accurate, consistent, well-annotated data, delivered efficiently.



***1xyq**: D.A. Lysek, C. Schorn, L.G. Nivon, V. Esteve-Moya, B. Christen, L. Calzolai, C. von Schroetter, F. Fiorito, T. Herrmann, P. Guntert, K. Wüthrich (2005) Prion protein NMR structures of cats, dogs, pigs, and sheep* Proc.Natl.Acad.Sci.USA ***102**:640-645.*

About the RCSB PDB

Photo credit: Wolfgang Bluhm

The RCSB PDB and some wwPDB partners

The RCSB PDB Organization

The RCSB PDB member institutions Rutgers, The State University of New Jersey and the San Diego Supercomputer Center and the Skaggs School of Pharmacy and Pharmaceutical Sciences at the University of California, San Diego (UCSD) jointly manage the project.

Photo credit: Wolfgang Bluhm

The RCSB PDB Advisory Board

Helen M. Berman, Director of the RCSB PDB, is a Board of Governors professor of chemical biology at Rutgers. Berman was part of the Brookhaven team that first envisioned the PDB. In addition, Berman is co-founder of the Nucleic Acid Database. Co-director Philip E. Bourne is a professor of pharmacology at UCSD and an adjunct professor at both The Burnham Institute and The Keck Graduate Institute.

The PDB also receives input from an international advisory board, made up of experts in X-ray crystallography, NMR, 3-D EM, bioinformatics, and education. The current membership includes: Edward N. Baker (Professor, University of Auckland), Manju Bansal (Professor, Indian Institute of Science), Stephen K. Burley (PDBAC Chair, Chief Scientific Officer and Senior Vice President, Research, Structural GenomiX), Wah Chiu (Professor, Baylor College of Medicine), Paul Craig (Professor, Rochester Institute of Technology), 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).

Funding

The RCSB PDB is supported by funds from the National Science Foundation (NSF), the National Institute of General Medical Sciences (NIGMS), the Office of Science, Department of Energy (DOE), the National Library of Medicine (NLM), the National Cancer Institute (NCI), the National Center for Research Resources (NCRR), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the National Institute of Neurological Disorders and Stroke (NINDS), and the National Institute of Diabetes & Digestive & Kidney Diseases (NIDDK).

NSF NIGMS Office of Science U.S. DEPARTMENT OF ENERGY NATIONAL LIBRARY OF MEDICINE NATIONAL CANCER INSTITUTE National Center for Research Resources NIBIB NATIONAL INSTITUTE OF NEUROLOGICAL DISORDERS AND STROKE NIDDK

Remediation

The wwPDB and the Remediation of the PDB Archive

A major achievement of the wwPDB has been the release of the remediated PDB archive (**remediation.wwpdb.org**).

Over the years, the wwPDB members have worked together to ensure the uniformity of entries archived in the PDB. The entire archive has now been reviewed, remediated, and publicly released at **ftp://ftp.wwpdb.org**. Files currently processed and released by the wwPDB sites reflect the new features incorporated into the archive as part of this project.

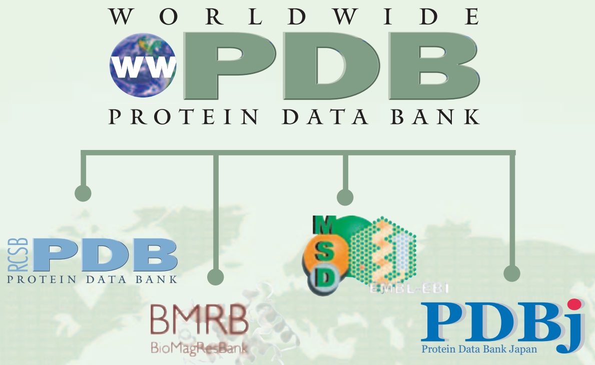
Remediated data are available for each PDB entry in three formats:

- mmCIF (**mmcif.pdb.org**). All remediation work was done using the PDB Exchange Dictionary (PDBx)[7] that follows the macromolecular Crystallographic Information File (mmCIF)[8] syntax.
- PDBML/XML (**pdbml.pdb.org**). Remediated data files are also available in PDBML/XML format,[9] in a direct translation from the files in mmCIF format.
- PDB File Format (**wwpdb.org**). The remediated files have been released in PDB File Format version 3.0. This version of the file format incorporates standardized atom nomenclature, and distinguishes deoxyribonucleic acid from ribonucleic acid.

The release of the remediated data has greatly improved the quality of searching and reporting possible across the PDB archive.

Remediation Highlights:

Sequence	Updated references to databases and taxonomies Resolved differences between chemical and macromolecular sequences
Citation	Verified and updated primary citation assignments
Assembly and virus information	Improved representation of deposited and experimental coordinate frames, symmetry, and frame transformations
Nucleic acid labeling	Deoxy and ribose nucleotides assigned separate chemical definitions
Beamline data	Beamline and synchrotron facility names have been made consistent with BioSync
Chemical components	Standardization of chemistry and nomenclature in monomers and ligands



About the wwPDB

The wwPDB consists of organizations that act as deposition, data processing and distribution centers for PDB data. The members are the RCSB PDB (USA), the Macromolecular Structure Database at the European Bioinformatics Institute (MSD-EBI), the Protein Data Bank Japan (PDBj) at Osaka University, and the BioMagResBank (BMRB) at the University of Wisconsin-Madison. The wwPDB's mission is to maintain a single Protein Data Bank archive of macromolecular structural data that is freely available to the global community.

Data deposited to the archive is processed using agreed-upon standards for full validation of the data. These data are forwarded to the RCSB PDB for release into the archive. wwPDB members also maintain websites that provide different views of the data.

Data Dictionaries: mmCIF and PDBx

The RCSB PDB uses mmCIF data dictionaries to describe the information content of PDB entries. The mmCIF dictionary is an ontology of more than 3,900 terms defining macromolecular structure and its related experiments.

The PDBx Dictionary consolidates content from a variety of dictionaries and includes extensions to describe NMR, EM, 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 PDBx 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.

The Chemical Component Dictionary

The Chemical Component Dictionary describes all small molecules and all standard and non-standard residues in the PDB archive. It has been remediated to address inconsistencies in older dictionary entries that resulted in valence problems, missing model coordinates, and redundant ligands.

The features of this new dictionary include standardized nomenclature; corrected model coordinates; and the addition of stereochemical assignments, aromatic bond assignments, idealized coordinates, chemical descriptors (SMILES & InChI)[5-6], and systematic chemical names.

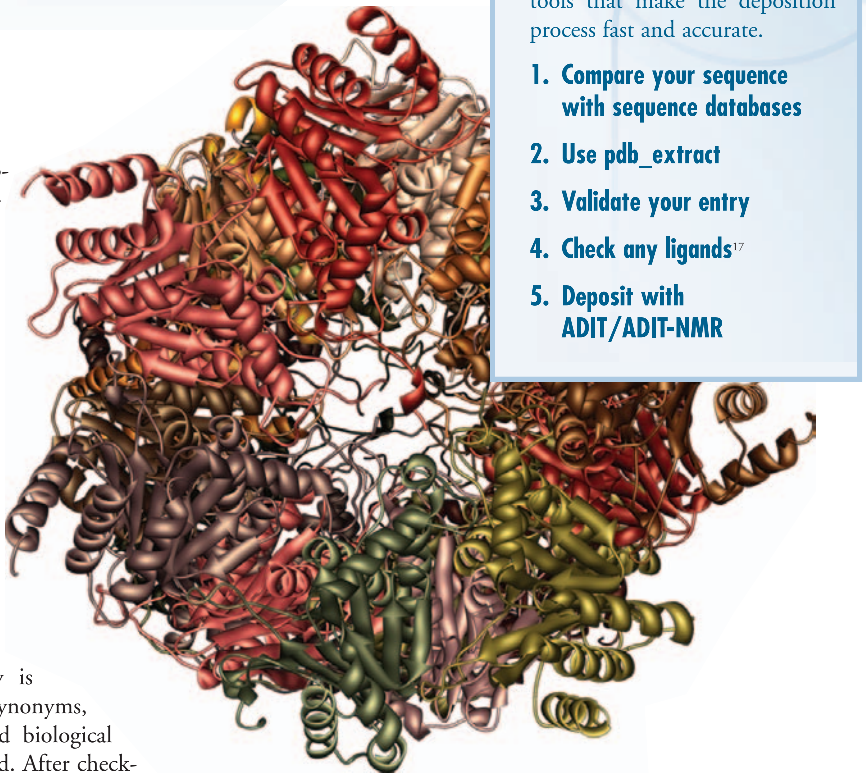
The full Chemical Component Dictionary and the companion Amino Acid Variants Dictionary is available for searching and downloading from **remediation.wwpdb.org/downloads.html**.

RCSB PDB Services: Data Deposition

5 Easy Steps for Structure Deposition

The RCSB PDB has developed tools that make the deposition process fast and accurate.

1. Compare your sequence with sequence databases
2. Use pdb_extract
3. Validate your entry
4. Check any ligands[17]
5. Deposit with ADIT/ADIT-NMR



1ryp: M. Groll, L. Ditzel, J. Lowe, D. Stock, M. Bochtler, H.D. Bartunik, R. Huber, (1997) Structure of 20S proteasome from yeast at 2.4 Å resolution. Nature ***386****:463-471.*

Data Input: Deposition, Validation, and Annotation

A key component of the wwPDB is the efficient capture (deposition) and curation (validation and annotation) of experimental structural data. Scientists contribute data produced from structure determination experiments using deposition tools available from the wwPDB partners. These data are then validated and annotated before being made publicly available. Data processed at the other wwPDB sites are forwarded to the RCSB PDB for inclusion in the archive.

When a structure is deposited, it is immediately assigned its own unique PDB ID.

Data annotators work to represent PDB data in the best possible way. Using tools developed by the RCSB PDB, data entries are carefully reviewed.

Annotators then compare the sequence reported in the deposition to sequence database references[10] and the citation reference with PubMed.[11] The entry is assigned a title, protein or other polymer names and synonyms, scientific name of the source of the protein(s) and biological assembly information. Any format errors are corrected. After checking the structure visually, annotators provide validation reports[12] and the completed coordinate file in mmCIF and PDB formats to the depositor for review. After corresponding with the depositor to finalize the entry for release, the complete entry, including its status information and PDB ID, is loaded into a relational database.

Depending upon the hold status selected by the depositor, data release occurs when a depositor gives approval to the annotated entry (status: REL), the hold date has expired (HOLD), or the journal article has been published (HPUB). Structures can be on HOLD or HPUB for no more than a year.

Software for Deposition and Validation

The RCSB PDB has developed a variety of tools to facilitate data annotation and validation by the depositor community.

pdb_extract[13] minimizes errors and saves time during the deposition process by extracting key details from the output files produced by many NMR and X-ray crystallographic applications so less information needs to be entered manually. The program merges these data into macromolecular Crystallographic Information File (mmCIF) format files that can be edited to add any additional information, and can be used with RCSB PDB tools for validation and deposition. pdb_extract can be used via a web interface or a workstation version that can be downloaded from pdb-extract.rcsb.org.

The RCSB PDB Validation Suite checks the format of the coordinate file and validates the overall structure before deposition. It produces a validation report containing geometrical and experimental checks from the programs MolProbity,[14] NUCheck, PROCHECK,[15] and SFCHECK.[16] Sequence/coordinate alignment, missing and extra atoms or residues, and data inconsistencies are also reported. The Validation Suite is also available for researchers to review the quality of any released structure before using it in their own research.

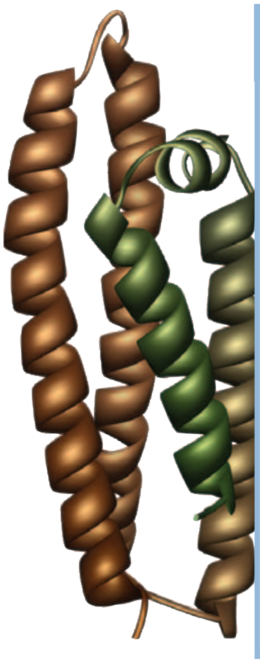
The deposition tool ADIT (AutoDep Input Tool) is available online from the RCSB PDB and PDBj, and is also available as a software download for standalone desktop use. ADIT provides access to a collection of programs for data input, validation, annotation, and format exchange. The ADIT system uses PDBx, which is based on the mmCIF dictionary.

The recently released ADIT-NMR lets depositors submit NMR structure and experimental data using a single tool. Available online from PDBj and BMRB, ADIT-NMR can be used to precheck, validate, and deposit NMR structures. Coordinates and restraint data are processed and released by the RCSB PDB and PDBj, while other NMR spectral data (such as chemical shifts, coupling constants, and relaxation parameters) are processed and archived by the BMRB.

RCSB PDB Services: Data Deposition

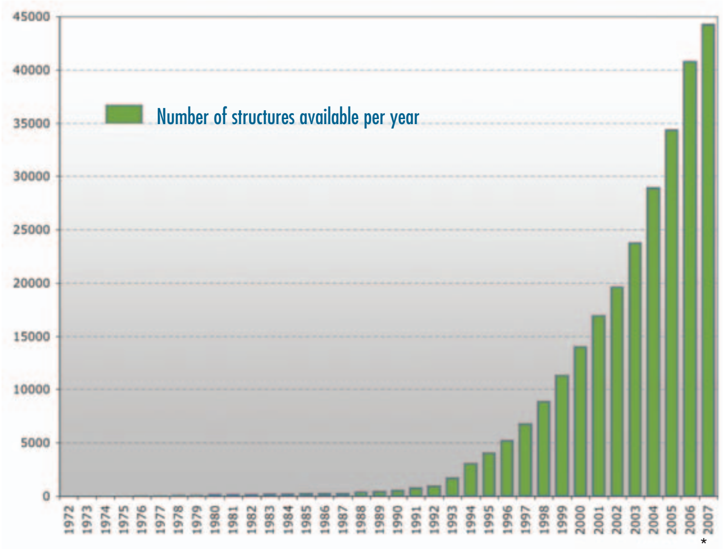
Statistics

During the period covered by this report, 8269 files were deposited to the wwPDB from around the world. More than half were processed by the RCSB PDB; structures were also processed by wwPDB members MSD-EBI and PDBj. Of these structures, approximately 85 percent were deposited with experimental data. Sequence data for about 55 percent of the depositions were released prior to the structure's release.

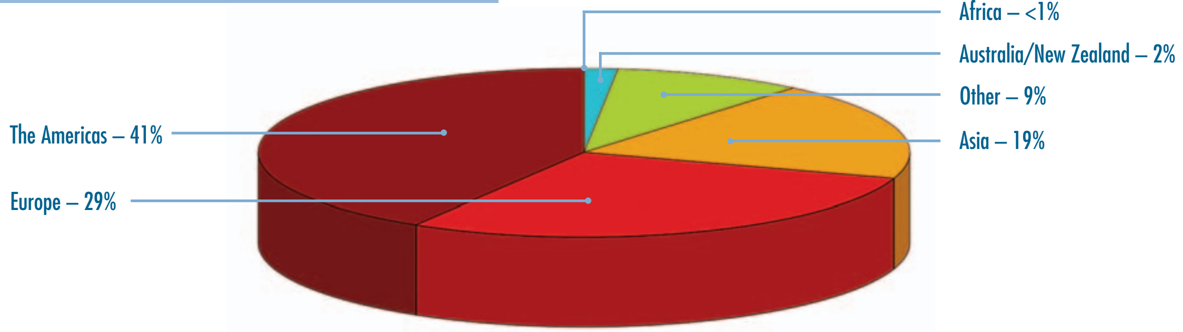


***1nfn:** L.M. Dong, S. Parkin, S.D. Trakhanov, B. Rupp, T. Simmons, K.S. Arnold, Y.M. Newhouse, T.L. Innerarity, K.H. Weisgraber (1996) Novel mechanism for defective receptor binding of apolipoprotein E2 in type III hyperlipoproteinemia.*
Nat.Struct.Biol. *3:718-722.*

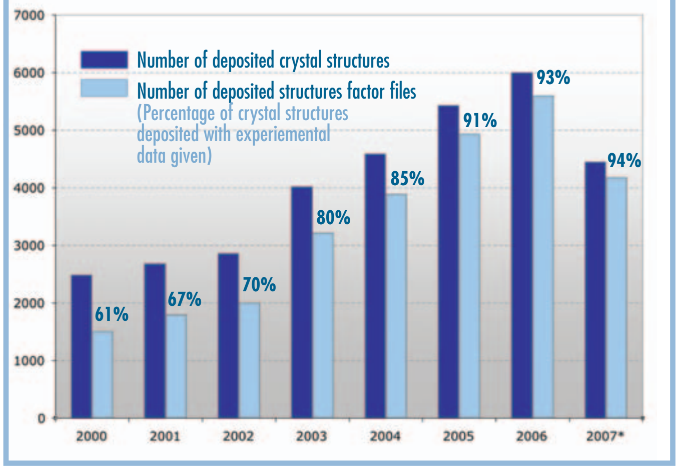
Growth of the Number of Structures Available in the PDB Archive



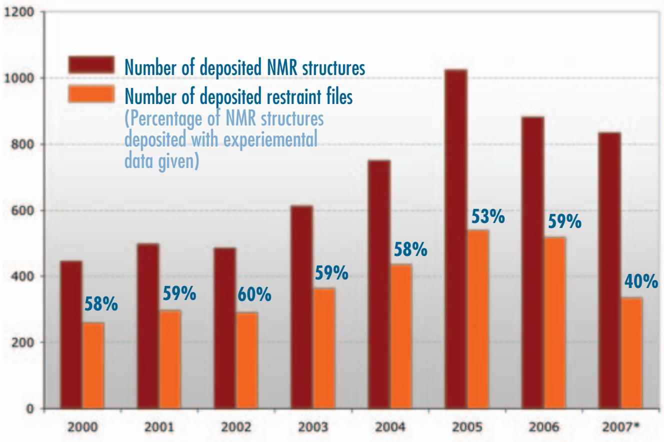
Depositions since 1997 by Depositor Location



Deposited Crystal Structures and Structures Factor Files



Deposited NMR Structures and Restraint Files



* only partial data shown for 2007

RCSB PDB Services: Data Query, Reporting & Access

Data Distribution and Access

RCSB PDB services and data are freely available through the internet.

The RCSB PDB website at www.pdb.org is accessed by about 100,000 unique visitors per month from nearly 140 different countries. Around 500 GigaBytes of data are transferred each month. On a typical weekday, two pages from the site are viewed every second. Data are accessed via the website, ftp server (supporting ftp and rsync access), Web Services and RSS feeds. The website is maintained 24 hours a day, seven days a week. A failover system automatically redirects internet traffic to a mirror site, if needed.

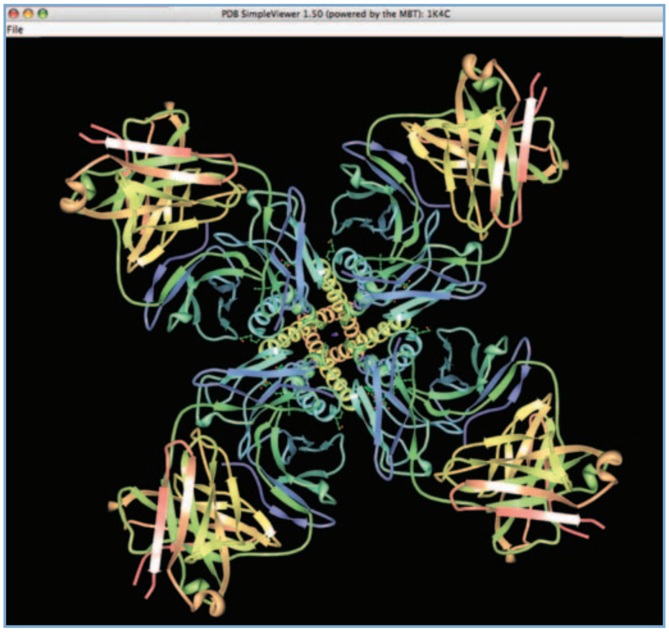
As the wwPDB archive keeper, the RCSB PDB updates the PDB archive at ftp://ftp.wwpdb.org weekly. The structures included in each release are highlighted on the RCSB PDB home page and clearly defined on the FTP site. During this report period, 7,324 coordinate files were released into the archive, along with 5,936 structure factor files, and 522 NMR restraint files.

Data Query and Reporting

The website is designed to make every structure accessible and comprehensible to all of our data users – students and teachers, biologists, structural biologists, computational biologists, and the general public.

More than 1000 curated web pages are available. The website also allows users to *search* for structures using simple and advanced search tools, and to *browse* the archive based on Gene Ontology (GO) terms, Enzyme Classification, Medical Subject Headings, Source Organism, Genome Location, and SCOP and CATH classifications.[11, 18-21]

Whether looking at individual or multiple structures, users have a variety of options for learning about and visualizing the entries. Each individual entry has a *Structure Summary* page that provides summary information, sequence details, static and interactive images of the molecule, and related links to other resources. A PDB, mmCIF or PDBML/XML format file for any structure can be *downloaded* as plain text or in one of several compressed formats. 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. All of these features are supported with help pages, Flash tutorials, and an active help desk.

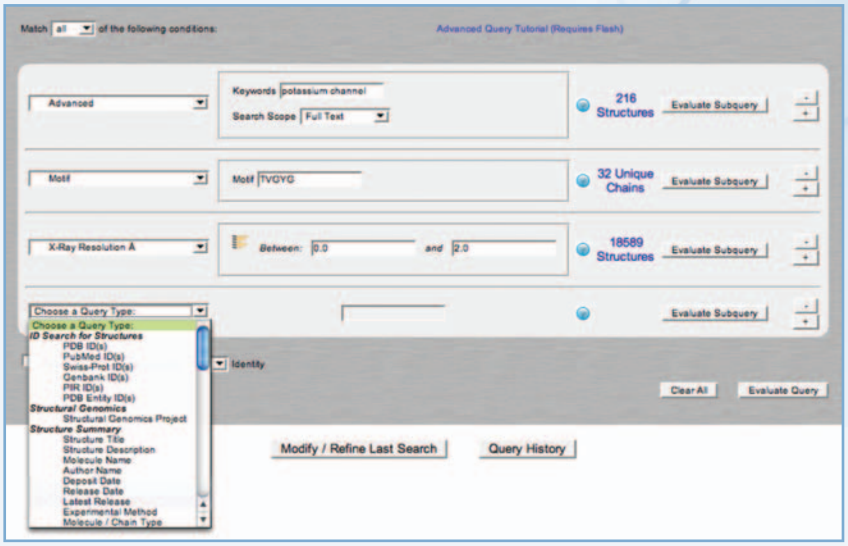


Simple Viewer can be used to visualize biological units. Shown: 1k4c (Y. Zhou, J.H. Morais-Cabral, A. Kaufman, and R. MacKinnon (2001) Chemistry of ion coordination and hydration revealed by a K+ channel-Fab complex at 2.0 Å resolution. Nature *414:43-8).*

Advanced Search

The data in the PDB archive offers a wealth of valuable metadata. *Advanced Search* is a powerful and easy-to-use interface to the database's powerful underlying search architecture and remediated data. Complex queries are constructed by combining simple "subqueries" chosen from a drop-down list. Users get a feel for the likely success of their search strategy while constructing the search by checking the number of results for each subquery.

A broad range of subqueries is available including sequence searches; GO assignments; SCOP and CATH domain assignments; and author name searches.



Match all of the following conditions:

Advanced — Keywords: potassium channel; Search Scope: Full Text — 216 Structures

Motif — Motif: TVGYG — 32 Unique Chains

X-Ray Resolution Å — Between: 0.0 and 2.0 — 18589 Structures

Choose a Query Type:

An example of an Advanced Search *which combines searches by keyword, sequence motif, and resolution. Additional queries can be added.*

RCSB PDB Services: Data Query, Reporting & Access

Utilizing Remediated Data

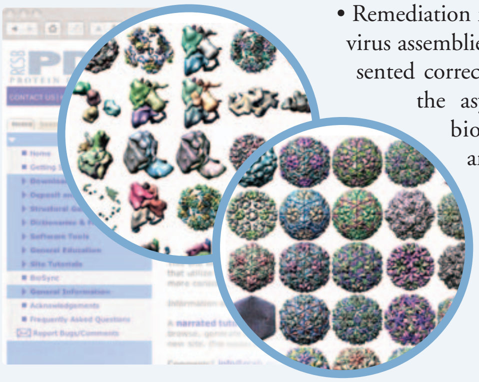
The center of the website is a database[22] that utilizes the data from the wwPDB Remediation Project. One of the goals of this project–the ability to search uniformly across the archive – is now possible. Since the underlying data files now describe PDB structures more accurately, queries can return a more accurate set of results.

Keyword or *Advanced Searches* offer different ways of exploring search results. Many options are available from the tabs shown above the default results list:

- New tabs provide the structures that map to related Gene Ontology (GO), SCOP and CATH results[11, 20-21]. Entries are returned in a tree browser, which indicates where these structures reside in the respective hierarchies. The SCOP tab, for example, indicates which hits belong to which class of proteins.

- A new *Sequence Details* report, available for each structure in the database, provides an exact mapping of structure sequence to the UniProt[23] sequence made possible through the Remediation Project. Users can select which properties, including domain definitions and secondary structure, are mapped to the sequence in this report.

- The primary citations for all structures have been verified as part of the Remediation Project. This improved mapping between structure and associated reference is reflected in the database. The *Citations Tab* included with the results of a search provides a PubMed-like list of the primary citations for the structures that match a query.



- Remediation means that full virus assemblies are now presented correctly. Images for the asymmetric and biological units are available for each structure and can be viewed as a "collage" for any search results list.

- The download option on the left menu provides access to either remediated or unremediated data in a variety of formats. If available, a report indicates how chain-naming conventions have changed to be more consistent.

- A ligand name can be entered in the keyword text search at the top of any page from the RCSB PDB website. The Advanced Search query engine can also be used to create searches for structures using ligands, based upon the ligand's name, ID code, or SMILES string. In addition to reviewing the structures that match the given query constraints, users can select the *Ligand Hits* tab, which lists the ligands known to interact with the structures matching the query. For example a keyword search for "protein kinase" will return all ligands known to bind protein kinases. The *Ligand Hits* tab also offers a gallery view of ligand images.

FTP: Downloading Tools and Time-stamped Copies

As part of a wwPDB initiative, time-stamped snapshots of the PDB archive are added each year to ftp://snapshots.rcsb.org. In addition to the yearly snapshot added in January 2007, a copy of the archive prior to the release of the remediated data is provided at ftp://ftp.rcsb.org. It is hoped that these snapshots will provide readily identifiable data sets for research on the PDB archives.

Scripts are available to help users download any of these ftp sites: the PDB archive (ftp.wwpdb.org), the frozen unremediated archive (ftp.rcsb.org), or any of the other time-stamped copies (snapshots.rcsb.org). These tools help users create local copies of all or part of the PDB archive.

Related Resources

A Portal for Structural Genomics

The international structural genomics efforts are focused on determining a large number of structures in a high throughput mode. Since the PDB is the repository for these protein structures, an emphasis is made toward developing resources that facilitate ways to deposit, track, and access these data. The RCSB PDB is actively involved in developing the informatics infrastructure needed for these projects, including the maintenance of data dictionaries describing these experiments. An online information portal at **sg.pdb.org** also offers online tools, summary reports, and target information related to structural genomics.[24]

Current and readily available information about the progress of protein production and structure solution at the structural genomics centers is tracked in databases at this site. The Target Registration Database (TargetDB; **targetdb.pdb.org**)[25] contains information about the progress of the production and solution of structures. The Protein Expression Purification and Crystallization Database (PepcDB; **pepcdb.pdb.org**)[24] extends the content of TargetDB with status history, stop conditions, reusable text protocols and contact information collected from the Protein Structure Initiative Centers funded by the NIGMS. These resources facilitate the coordination needed between the different centers to promote efficient structure solution.

The portal links to sites that offer analyses of functional annotation, including a tool from the RCSB PDB that can be used to explore the distributions of functions found among structural genomics structures, PDB structures, genomes, and homology models according to enzyme classification, GO terms, and disease.[26]

Cryo-EM Definition Development Project

The RCSB PDB and the MSD-EBI are working together to establish a "one-stop shop" for deposition and retrieval of cryo-EM data. The goal is to create a robust system for the collection, validation, annotation, distribution and visualization of structures derived from cryo-EM experiments and analyses.

A comprehensive dictionary of cryo-EM data items is being developed. The present dictionary (at **mmcif.pdb.org**) includes 54 categories and more than 500 data items to describe different aspects of single particle, 2D electron diffraction, and helical diffraction cryo-EM experiments.

Community Interactions

The RCSB PDB has a diverse user community of depositors, data users, students, teachers, and the general public. Through exhibits, presentations, and an active help desk, we gain feedback for further development of the RCSB PDB while providing materials that promote scientific literacy and a broader understanding of structural biology.

Publications

www.pdb.org is updated weekly with news, recent developments, new resources, and improvements to existing documents. Educational features, such as new *Molecule of the Month* installments, are posted regularly.

Published quarterly, the RCSB PDB Newsletter describes and highlights recent RCSB PDB activities. Features include an Education Corner that describes how the PDB archive and RCSB PDB resources are used in the classroom, and a Community Focus interview with many of the scientific luminaries in the PDB user community.

A variety of flyers, brochures, and tutorials are distributed to users and published online.

The RCSB PDB regularly contributes articles to many peer-reviewed journals.[27-32]

PDB NEWSLETTER
Summer 2007 • Number 34
POLIOVIRUS AND
5 Easy Steps for Structure Deposition
Structural Biologist's Guide to Synchrotron Facilities
BioSync Web Resource
BioSync.rcsb.org
BioSync@deposit.rcsb.org
www.pdb.org • info@rcsb.org
www.pdb.org • deposit@deposit.rcsb.org
RESEARCH COLLABORATORY FOR STRUCTURAL BIOINFORMATICS

2bk1: S.J. Tilley, E.V. Orlova, R.J. Gilbert, P.W. Andrew, and H.R. Saibil (2005) Structural basis of pore formation by the bacterial toxin pneumolysin. Cell *121:247-56.*

BioSync

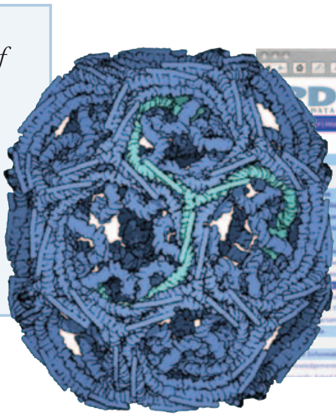
BioSync (Structural Biology Synchrotron Users Organization) is an online clearinghouse for beamline information at synchrotron facilities.[33] The website contains descriptions of all operational synchrotron beamlines currently being used for single crystal macromolecular crystallography. The site also includes PDB deposition statistics with galleries of structures, cross-linked to the RCSB PDB, that are grouped by site and beamline. Statistics are updated weekly as new structures are released into the PDB archive (**biosync.pdb.org**).

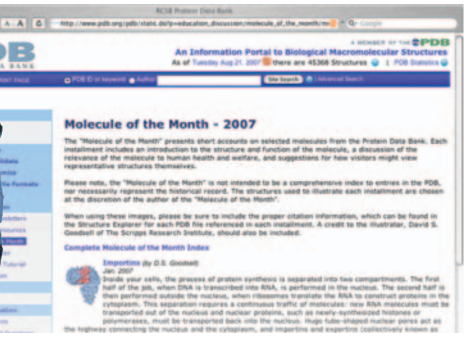
Educational Resources

Thousands of scientists visit **www.pdb.org** every day. However, advanced researchers are not the RCSB PDB's only users. The science of protein and nucleic acid structure is also available to students, educators, and the general public through a number of outreach programs and resources.

Accessible from the RCSB PDB homepage, the *Molecule of the Month* series describes selected molecules from the PDB archive. Each installment introduces the structure and function of a molecule, and relates it to human health and welfare.

1xi4: *A. Fotin, Y. Cheng, P. Sliz, N. Grigorieff, S.C. Harrison, T. Kirchhausen, T. Walz (2004) Molecular model for a complete clathrin lattice from electron cryomicroscopy.* Nature **432**:*573-579.*





RCSB Protein Data Bank

http://www.pdb.org/pdb/static.do?p=education_discussion/molecule_of_the_month/m

A MEMBER OF THE PDB

An Information Portal to Biological Macromolecular Structures

As of Tuesday Aug 21, 2007 there are 45368 Structures | PDB Statistics

PRINT PAGE ○ PDB ID or keyword ● Author [Site Search] | Advanced Search

Molecule of the Month - 2007

The "Molecule of the Month" presents short accounts on selected molecules from the Protein Data Bank. Each installment includes an introduction to the structure and function of the molecule, a discussion of the relevance of the molecule to human health and welfare, and suggestions for how visitors might view representative structures themselves.

Please note, the "Molecule of the Month" is not intended to be a comprehensive index to entries in the PDB, nor necessarily represent the historical record. The structures used to illustrate each installment are chosen at the discretion of the author of the "Molecule of the Month".

When using these images, please be sure to include the proper citation information, which can be found in the Structure Explorer for each PDB file referenced in each installment. A credit to the illustrator, David S. Goodsell of The Scripps Research Institute, should also be included.

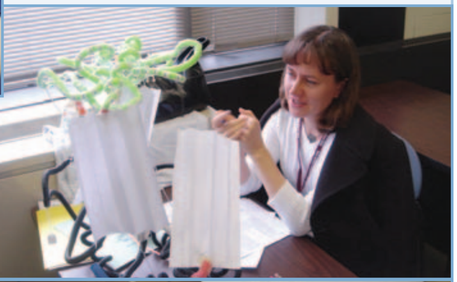
Complete Molecule of the Month Index

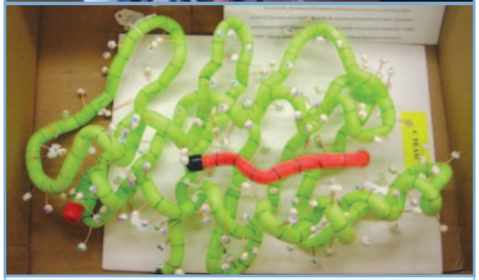
Importins (by D.S. Goodsell)
Jan. 2007
Inside your cells, the process of protein synthesis is separated into two compartments. The first half of the job, when DNA is transcribed into RNA, is performed in the nucleus. The second half is then performed outside the nucleus, when ribosomes translate the RNA to construct proteins in the cytoplasm. This separation requires a continuous traffic of molecules: new RNA molecules must be transported out of the nucleus and nuclear proteins, such as newly-synthesized histones or polymerases, must be transported back into the nucleus. Huge tube-shaped nuclear pores act as the highway connecting the nucleus and the cytoplasm, and importins and exportins (collectively known as

Interactive and automatic displays have been created to showcase protein structures. One animation that provides information about specific proteins found in marine organisms–and in the PDB archive–was part of the *Sea of Genes* exhibit at the Birch Aquarium (Scripps Institution of Oceanography at University of California, San Diego).

Princeton High School won the 2007 New Jersey Science Olympiad Protein Modeling event with their model of a Major Histocompatibility Complex.

In this trial event, teams of high school students use the resources of the RCSB PDB to build a 3D model of an assigned protein.





The RCSB PDB exhibits at education-related meetings to talk to teachers and students about protein structure and function.

Teachers also become students of the RCSB PDB at programs targeted at educators, such as TeacherTECH at the San Diego Supercomputer Center.

In New Jersey, the RCSB PDB supplies the kits* used to create the models, judges the proteins, and hosts a Coaches Workshop.

* Mini-Toobers are products of 3D Molecular Designs
www.3dmoleculardesigns.com

The 3D nature of proteins is engaging and draws the attention of students of all ages. It can also lead to interesting hands-on activities.

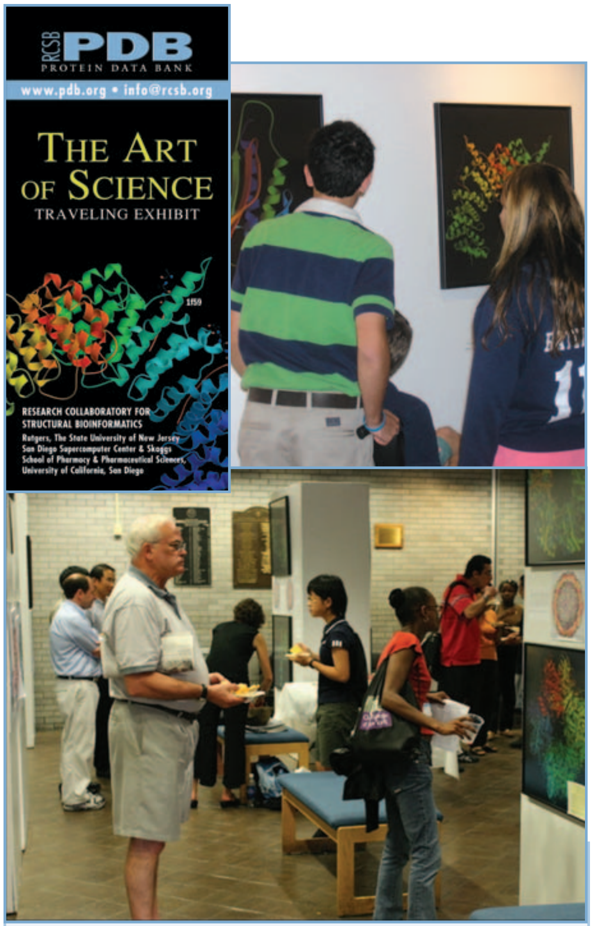
Other resources include:

- The General Education section of the RCSB PDB website offers a variety of resources for teachers and students.
- The electronic help desk at **info@rcsb.org** provides around-the-clock support for using RCSB PDB resources and beyond.

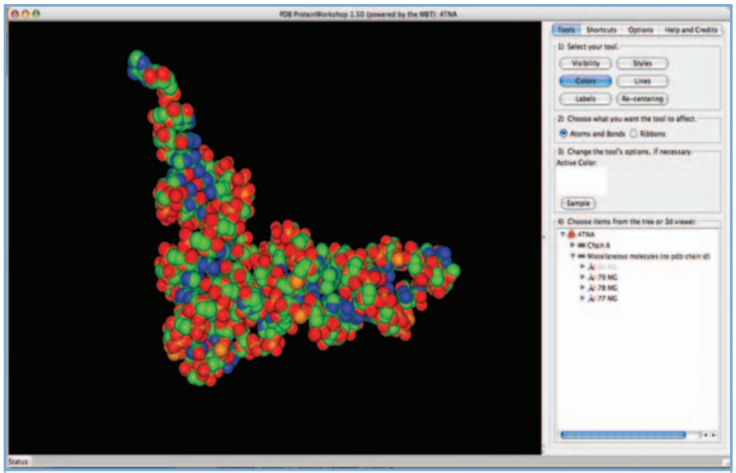
A poster prize is also awarded at professional society meetings to recognize outstanding student achievement in structural biology.

Students and teachers visit the RCSB PDB on field trips and as part of educational programs.

RCSB PDB members also utilize the site when teaching graduate and undergraduate college courses.


RCSB PDB
PROTEIN DATA BANK
www.pdb.org • info@rcsb.org
THE ART OF SCIENCE
TRAVELING EXHIBIT
1I59
RESEARCH COLLABORATORY FOR STRUCTURAL BIOINFORMATICS
Rutgers, The State University of New Jersey
San Diego Supercomputer Center & Skaggs School of Pharmacy & Pharmaceutical Sciences, University of California, San Diego

The traveling *Art of Science* exhibit showcases the beauty and variety of protein shapes found in the RCSB PDB, bringing structural biology to art enthusiasts from New York to India.


PDB ProteinWorkshop 1.50 (powered by the MBT): 4TNA
Tools
Shortcuts
Options
Help and Credits
1) Select your tool.
Visibility
Styles
Colors
Lines
Labels
Re-centering
2) Choose what you want the tool to affect.
● Atoms and Bonds ○ Ribbons
3) Change the tool's options, if necessary.
Active Color:
Sample
4) Choose items from the tree or 3d viewer.
4TNA
Chain A
Miscellaneous molecules (no pdb chain id)
79 MG
78 MG
77 MG
Status

A variety of tools are available to visualize structure. Using the MBT tools SimpleViewer and Protein Workshop,[34] users can easily view structures and create images.

*Shown in Protein Workshop: **4tna**. (B. Hingerty, R.S. Brown, A. Jack (1978) Further refinement of the structure of yeast tRNAPhe.* J.Mol.Biol. ***124**:523-534.)*

References

Cover Structures Citations

1a6c: V. Chandrasekar and J.E. Johnson (1998) The structure of tobacco ringspot virus: a link in the evolution of icosahedral capsids in the picornavirus superfamily. *Structure* **6**:157-171.

1ayn: M.A. Oliveira, R. Zhao, W.M. Lee, M.J. Kremer, I. Minor, R.R. Rueckert, G.D. Diana, D.C. Pevear, F.J. Dutko, M.A. McKinlay, M.G. Rossmann (1993) The structure of human rhinovirus 16. *Structure* **1**:51-68.

1dzl: X.S. Chen, R.L. Garcea, I. Goldberg, G. Casini, S.C. Harrison (2000) Structure of small virus-like particles assembled from the L1 protein of human papillomavirus 16. **Mol.Cell** **5**:557-567.

1ej6: K.M. Reinisch, M.L. Nibert, S.C. Harrison (2000) Structure of the reovirus core at 3.6 Å resolution. *Nature* **404**:960-967.

1f2n: C. Qu, L. Liljas, N. Opalka, C. Brugidou, M. Yeager, R.N. Beachy, C.M. Fauquet, J.E. Johnson, T. Lin (2000) 3D domain swapping modulates the stability of members of an icosahedral virus group. *Structure Fold.Des.* **8**:1095-1103.

1gw8: C. San Martin, J.T. Huiskonen, J.K. Bamford, S.J. Butcher, S.D. Fuller, D.H. Bamford, R.M. Burnett (2002) Minor proteins, mobile arms and membrane-capsid interactions in the bacteriophage PRD1 capsid. *Nat.Struct.Biol.* **9**:756-763.

1ihm: B.V. Prasad, M.E. Hardy, T. Dokland, J. Bella, M.G. Rossmann, M.K. Estes (1999) X-ray crystallographic structure of the Norwalk virus capsid. *Science* **286**:287-290.

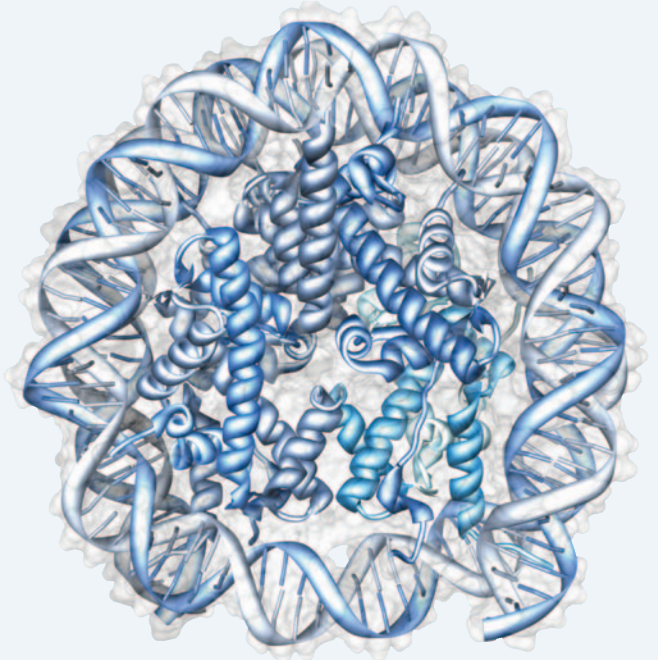
1k4r: R.J. Kuhn, W. Zhang, M.G. Rossmann, S.V. Pletnev, J. Corver, E. Lenches, C.T. Jones, S. Mukhopadhyay, P.R. Chipman, E.G. Strauss, T.S. Baker, J.H. Strauss (2002) Structure of dengue virus: implications for flavivirus organization, maturation, and fusion. *Cell* **108**:717- 725.

1m1c: H. Naitow, J. Tang, M. Canady, R.B. Wickner, J.E. Johnson (2002) L-A virus at 3.4 Å resolution reveals particle architecture and mRNA decapping mechanism. *Nat.Struct.Biol.* **9**:725-728.

1mvb: S.H. Van den Worm, N.J. Stonehouse, K. Valegard, J.B. Murray, C. Walton, K. Fridborg, P.G. Stockley, L. Liljas (1998) Crystal structures of MS2 coat protein mutants in complex with wild-type RNA operator fragments. *Nucleic Acids Res.* **26**:1345-1351.

1rug: A.T. Hadfield, M.A. Oliveira, K.H. Kim, I. Minor, M.J. Kremer, B.A. Heinz, D. Shepard, D.C. Pevear, R.R. Rueckert, M.G. Rossmann (1995) Structural studies on human rhinovirus 14 drug-resistant compensation mutants. *J.Mol.Biol.* **253**:61-73.

1qju: A.T. Hadfield, G.D. Diana, M.G. Rossmann (1999) Analysis of three structurally related anti-viral compounds in complex with human rhinovirus 16. *Proc.Natl.Acad.Sci.USA* **96**:14730-14735.



1kx3: C.A. Davey, D.F.Sargent, K. Luger, A.W. Maeder, T.J. Richmond (2002) Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 Å resolution J.Mol.Biol. *319:1097-1113.*

1tnv: M. Bando, Y. Morimoto, T. Sato, T. Tsukihara (1994) Crystal structural analysis of tobacco necrosis virus at 5 Å resolution. *Acta Crystallogr.,Sect.D* **50**:878-883.

1upn: D. Bhella, I.G. Goodfellow, P. Roversi, D. Pettigrew, Y. Chaudhry, D.J. Evans, S.M. Lea (2004) The structure of echovirus type 12 bound to a two-domain fragment of its cellular attachment protein decay-accelerating factor (CD 55). *J.Biol.Chem.* **279**:8325-8332.

1wcd: F. Coulibaly, C. Chevalier, I. Gutsche, J. Pous, J. Navaza, S. Bressanelli, B. Delmas, F.A. Rey (2005) The birnavirus crystal structure reveals structural relationships among icosahedral viruses. *Cell* **120**:761-772.

1vak: V. Sangita, G.L. Lokesh, P.S. Satheshkumar, C.S. Vijay, V. Saravanan, H.S. Savithri, M.R. Murthy (2004) T=1 capsid structures of Sesbania mosaic virus coat protein mutants: determinants of T=3 and T=1 capsid assembly. *J.Mol.Biol.* **342**:987-999.

1yc6: S.B. Larson, R.W. Lucas, A. McPherson (2005) Crystallographic structure of the T=1 particle of brome mosaic virus. *J.Mol.Biol.* **346**:815-831.

2btv: J.M. Grimes, J.N. Burroughs, P. Gouet, J.M. Diprose, R. Malby, S. Zientara, P.P. Mertens, D.I. Stuart (1998) The atomic structure of the bluetongue virus core. *Nature* **395**:470-478.

2buk: T.A. Jones and L. Liljas (1984) Structure of satellite tobacco necrosis virus after crystallographic refinement at 2.5 Å resolution. *J.Mol.Biol.* **177**:735-767.

2c6s: C. Zubieta, L. Blanchoin, S. Cusack (2006) Structural and biochemical characterization of a human adenovirus 2/12 penton base chimera. *FEBS J.* **273**:4336-4345.

2fte: L. Gan, J.A. Speir, J.F. Conway, G. Lander, N. Cheng, B.A. Firek, R.W. Hendrix, R.L. Duda, L. Liljas, J.E. Johnson (2006) Capsid conformational sampling in HK97 maturation visualized by X-ray crystallography and cryo-EM. *Structure* **14**:1655-1665.

2gh8: R. Chen, J.D. Neill, M.K. Estes, B.V. Prasad (2006) X-ray structure of a native calicivirus: Structural insights into antigenic diversity and host specificity. *Proc.Natl.Acad.Sci.USA* **103**:8048-8053.

Text References

1. Protein Data Bank (1971) Protein Data Bank. *Nature New Biol.* **233**:223.
2. F.C. Bernstein, T.F. Koetzle, G.J.B. Williams, E.F. Meyer Jr., M.D. Brice, J.R. Rodgers, O. Kennard, T. Shimanouchi, M. Tasumi (1977) Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**:535-542.
3. H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne (2000) The Protein Data Bank. *Nucleic Acids Res.* **28**:235-242.
4. H.M. Berman, K. Henrick, H. Nakamura (2003) Announcing the worldwide Protein Data Bank. *Nat Struct Biol.* **10**:980.
5. D. Weininger (1988) SMILES 1. Introduction and encoding rules. *J. Chem. Inf. Comput. Sci.* **28**:31.
6. © The International Union of Pure and Applied Chemistry, IUPAC International Chemical Identifier (InChI). (2005) contact: secretariat@iupac.org.
7. J. Westbrook, K. Henrick, E.L. Ulrich, H.M. Berman, 3.6.2 The Protein Data Bank exchange data dictionary, in *International Tables for Crystallography, G. Definition and exchange of crystallographic data*, S.R. Hall and B. McMahon, Editors. 2005, Springer: Dordrecht, The Netherlands. p. 195-198.
8. P.M.D. Fitzgerald, J.D. Westbrook, P.E. Bourne, B. McMahon, K.D. Watenpaugh, H.M. Berman, 4.5 Macromolecular dictionary (mmCIF), in *International Tables for Crystallography, G. Definition and exchange of crystallographic data*, S.R. Hall and B. McMahon, Editors. 2005, Springer: Dordrecht, The Netherlands. p. 295-443.
9. J. Westbrook, N. Ito, H. Nakamura, K. Henrick, H.M. Berman (2005) PDBML: The representation of archival macromolecular structure data in XML. *Bioinformatics.* **21**:988-992.
10. S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**:403-410.
11. D.L. Wheeler, T. Barrett, D.A. Benson, S.H. Bryant, K. Canese, V. Chetvernin, D.M. Church, M. DiCuccio, R. Edgar, S. Federhen, L.Y. Geer, Y. Kapustin, O. Khovayko, D. Landsman, D.J. Lipman, T.L. Madden, D.R. Maglott, J. Ostell, V. Miller, K.D. Pruitt, G.D. Schuler, E. Sequeira, S.T. Sherry, K. Sirotkin, A. Souvorov, G. Starchenko, R.L. Tatusov, T.A. Tatusova, L. Wagner, E. Yaschenko (2007) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **35**(Database issue):D5-12.
12. J. Westbrook, Z. Feng, K. Burkhardt, H.M. Berman (2003) Validation of protein structures for the Protein Data Bank. *Meth Enz.* **374**:370-385.
13. H. Yang, V. Guranovic, S. Dutta, Z. Feng, H.M. Berman, J. Westbrook

(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.

14. I.W. Davis, L.W. Murray, J.S. Richardson, D.C. Richardson (2004) MOLPROBITY: structure validation and all-atom contact analysis for nucleic acids and their complexes. *Nucleic Acids Res.* **32**(Web Server issue):W615-9.
15. R.A. Laskowski, M.W. McArthur, D.S. Moss, J.M. Thornton (1993) PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.* **26**:283-291.
16. A.A. Vaguine, J. Richelle, S.J. Wodak (1999) SFCHECK: a unified set of procedures for evaluating the quality of macromolecular structure-factor data and their agreement with the atomic model. *Acta Crystallogr D Biol Crystallogr.* **55**:191-205.
17. Z. Feng, L. Chen, H. Maddula, O. Akcan, R. Oughtred, H.M. Berman, J. Westbrook (2004) Ligand Depot: A data warehouse for ligands bound to macromolecules. *Bioinformatics* **20**:2153-2155.
18. The Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. *Nature Genetics* **25**:25-29.
19. Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and classification of enzymes by the reactions they catalyse. **www.chem.qmw.ac.uk/iubmb/enzyme**.
20. L. Conte, A. Bart, T. Hubbard, S. Brenner, A. Murzin, and C. Chothia (2000) SCOP: a structural classification of proteins database. *Nucleic Acids Res.* **28**:257-259.
21. C.A. Orengo, A.D. Michie, S. Jones, D.T. Jones, M.B. Swindells, and J.M. Thornton (1997) CATH–a hierarchic classification of protein domain structures. *Structure* **5**:1093-1108.
22. N. Deshpande, K.J. Addess, W.F. Bluhm, J.C. Merino-Ott, W. Townsend-Merino, Q. Zhang, C. Knezevich, L. Xie, L. Chen, Z. Feng, R. Kramer Green, J.L. Flippen-Anderson, J. Westbrook, H.M. Berman, P.E. Bourne (2005) The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema. *Nucleic Acids Res.* **33**:D233-D237.
23. The UniProt Consortium (2007) The Universal Protein Resource (UniProt). *Nucleic Acids Res.* **35**(Database issue):D193-7.
24. A. Kouranov, L. Xie, J. de la Cruz, L. Chen, J. Westbrook, P.E. Bourne, H.M. Berman (2006) The RCSB PDB information portal for structural genomics. *Nucleic Acids Res.* **34**:D302-D305.
25. L. Chen, R. Oughtred, H.M. Berman, J. Westbrook (2004) TargetDB: A target registration database for structural genomics projects. *Bioinformatics* **20**:2860-2862.
26. L. Xie and P.E. Bourne (2005) Functional coverage of the human genome by existing structures, structural genomics targets, and homology models. *PLoS Comp Biol.* **1**:e31.
27. P. Bourne, W.F. Bluhm, N. Deshpande, Q. Zhang, H.M. Berman, J.L. Flippen-Anderson, The Research Collaboratory for Structural Bioinformatics Protein Data Bank, in *Comprehensive Medicinal Chemistry II*, **Vol. 3**, Drug Discovery Technologies, H. Kubinyi, Editor. 2007, Elsevier: Oxford, UK. p. 373-388.
28. H.M. Berman, K. Henrick, H. Nakamura, J.L. Markley (2007) The Worldwide Protein Data Bank (wwPDB): Ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.* **35**(Database issue):D301-3.
29. T.A. Holland, S. Veretnik, I.N. Shindyalov, P.E. Bourne (2006) Partitioning protein structures into domains: Why is it so difficult? *J Mol Biol.* **361**:562-90.
30. K. Burkhardt, B. Schneider, J. Ory (2006) A biocurator perspective: annotation at the Research Collaboratory for Structural Bioinformatics Protein Data Bank. *PLoS Comput Biol.* **2**:e99.
31. P. Bourne and J. McEntyre (2006) Biocurators: Contributors to the World of Science. *PLoS Comput Biol.* **2**:e142.
32. H.M. Berman, S.K. Burley, W. Chiu, A. Sali, A. Adzhubei, P.E. Bourne, S.H. Bryant, J. Roland L. Dunbrack, K. Fidelis, J. Frank, A. Godzik, K. Henrick, A. Joachimiak, B. Heymann, D. Jones, J.L. Markley, J. Moult, G.T. Montelione, C. Orengo, M.G. Rossmann, B. Rost, H. Saibil, T. Schwede, D.M. Standley, J.D. Westbrook (2006) Outcome of a workshop on archiving structural models of biological macromolecules. *Structure* **14**: 1211-1217.
33. A. Kuller, W. Fleri, W.F. Bluhm, J.L. Smith, J. Westbrook, P.E. Bourne (2002) A biologist's guide to synchrotron facilities: the BioSync web resource. *TIBS* **27**: 213-215.
34. J.L. Moreland, A. Gramada, O.V. Buzko, Q. Zhang, P.E. Bourne (2005) The Molecular Biology Toolkit (MBT): A modular platform for developing molecular visualization applications. *BMC Bioinformatics* **6**:21.
35. E.F. Pettersen, T.D. Goddard, C.C. Huang, G.S. Couch, D.M. Greenblatt, E.C. Meng, T.E. Ferrin (2004) UCSF Chimera–a visualization system for exploratory research and analysis. *J Comput Chem.* **25**:1605-12.

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

BioMagResBank
www.bmrb.wisc.edu

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 and the Skaggs School of
Pharmacy and Pharmaceutical
Sciences, University of
California, San Diego**
9500 Gilman Drive
La Jolla, CA 92093-0537

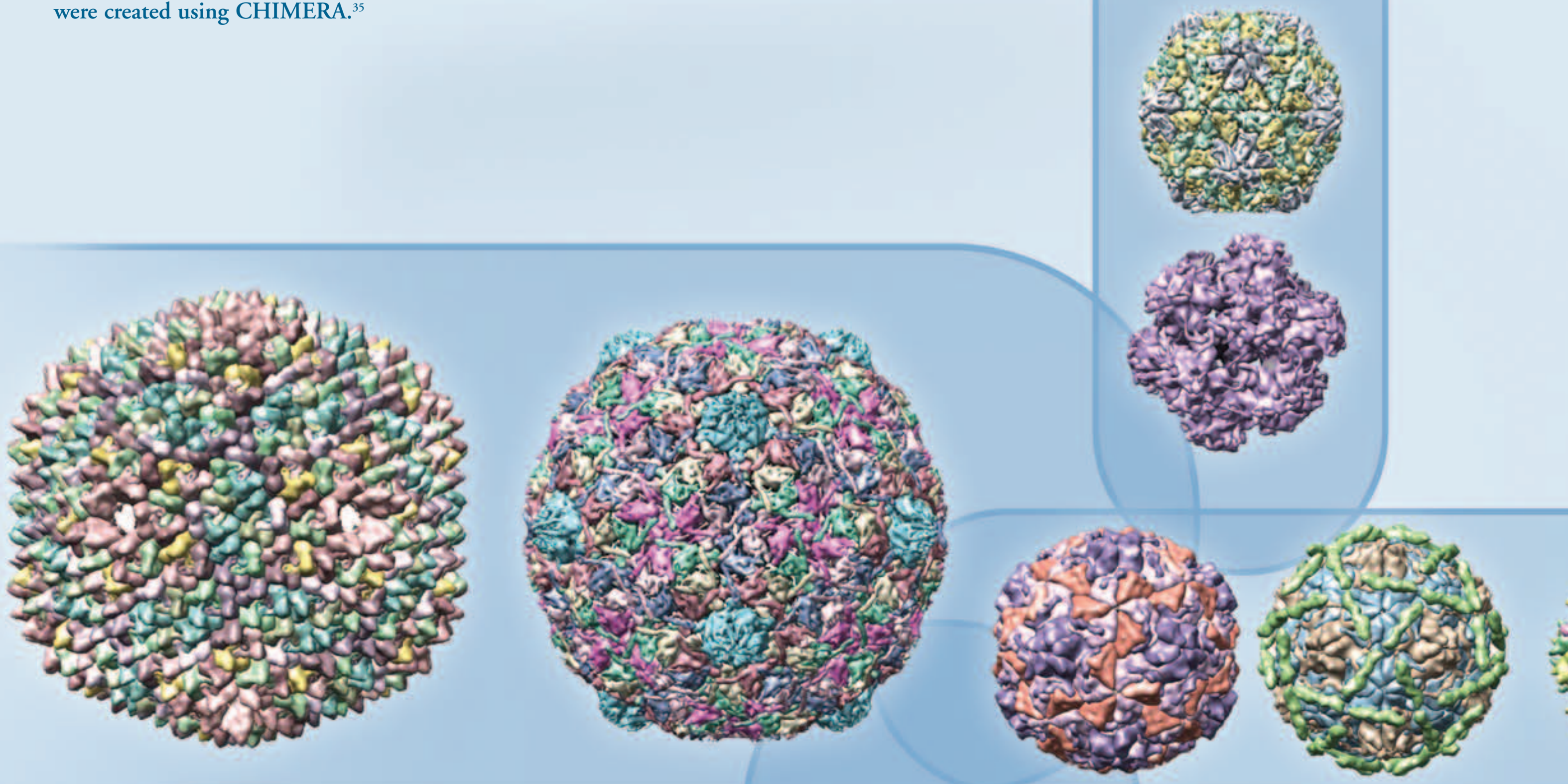
RCSB PDB Management

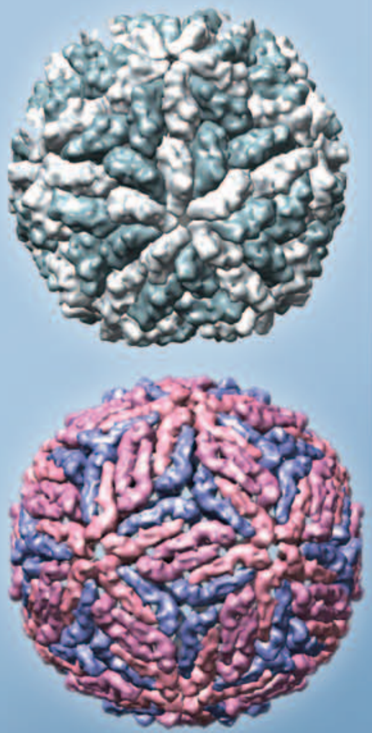
Dr. Helen M. Berman
Director
Board of Governors Professor of
Chemistry & Chemical Biology
Rutgers, The State University
of New Jersey
berman@rcsb.rutgers.edu

Dr. Philip E. Bourne
Co-Director
Professor of Pharmacology,
UCSD Adjunct Professor, The
Burnham Institute & The Keck
Graduate Institute
bourne@sdsc.edu

A list of current RCSB PDB Team Members is available at **www.pdb.org**.

Images of the biological
macromolecules in this report
were created using CHIMERA.[35]





RCSB PDB

www.pdb.org

Send questions or comments to:

info@rcsb.org