Fall 2009 • Number 43

RCSB PDB PROTEIN DATA BANK

Published quarterly by the Research Collaboratory for Structural Bioinformatics Protein Data Bank

NEWSLETTER

Contents

SNAPSHOT: OCTOBER 1, 2009

60,464 released atomic coordinate entries

MOLECULE TYPE		EXPERIMENTAL TECHNIQUE	
55,917	proteins, peptides, and viruses	52,017	X-ray
2,060	nucleic acids	8,043	NMR
2,454	protein/nucleic acid complexes	254	electron microscopy
33	other	18	hybrid
		132	other

41,210 structure factor files
5,329 NMR restraint files

Participating RCSB Members:
Rutgers • SDSC/SKAGGS/UCSD
E-mail: info@rcsb.org
Web: www.pdb.org • FTP: ftp.wwpdb.org

The RCSB PDB is a member of the wwPDB (www.wwpdb.org)

Weekly RCSB PDB news is available online at www.pdb.org

Message from the RCSB PDB: What's New?

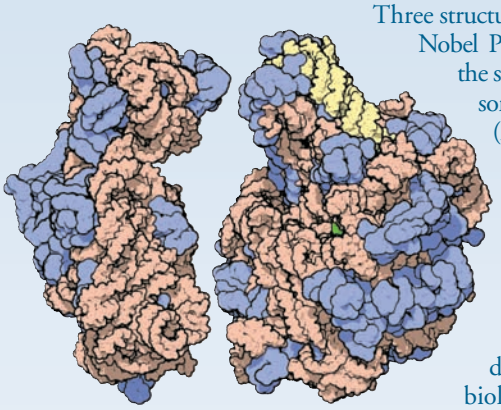
Each week, approximately 140 structures and related experimental data are released into the PDB archive and loaded into RCSB PDB's database.

To provide tools to query, report, and view these structures, the website at www.pdb.org is regularly updated with new features and resources. The WHAT'S NEW button at the top of the each web page links to a description of the latest tools and enhancements. Detailed instructions, examples, and in some cases, screencasts illustrate how to access and use these features. Recent examples include new options available for reviewing query results and the Comparison Tool for calculating pairwise sequence and structure alignments.

Detailed news about all RCSB PDB activities is published online each week, and in this quarterly newsletter.

We welcome your questions and comments about these new features.

Nobel Prize for the Ribosome



Three structural biologists have won the 2009 Nobel Prize in Chemistry for studies of the structure and function of the ribosome–Venkatraman Ramakrishnan (MRC Laboratory of Molecular Biology), Thomas A. Steitz (Yale University), and Ada E. Yonath (Weizmann Institute of Science). The depositions of their first complete ribosome subunit structures (PDB IDs 1fjg, 1ffk, and 1fka) almost a decade ago ushered structural biology into a new era.

Since that time, more than 120 ribosome structures consisting of 50S, 30S subunits and complete 70S ribosomes have been contributed by these Nobel scientists. The structures, complexed with and without antibiotics, tRNAs, mRNAs, initiation factors, and release factors, provide a basis for understanding how the ribosome works and are useful tools for drug development.

Ribosome-related resources at the RCSB PDB include the October 2000 edition of the *Molecule of the Month*, animated GIFs of the large and small ribosomal subunit, and a small poster commemorating the award.

Data Deposition and Processing

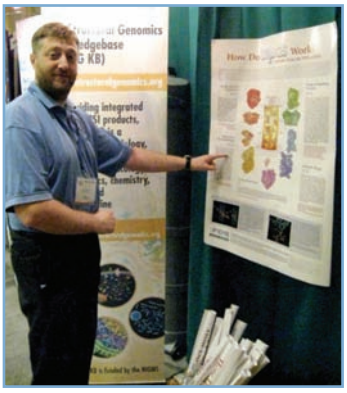
SF-Tool: A Tool for Crystallographic Experimental Data Validation

A streamlined, web-based tool can validate crystallographic experimental data. SF-Tool can be used to:

- Validate model coordinates against structure factor data (using SFCheck)[1]
- Easily translate your structure factor file between different formats (mmCIF, MTZ, CNS/CNX, XPLOR, SHELX, TNT, HKL2000, SCALEPACK, D*Trek, SAINT, or OTHER format)
- Check for twinned or detwinned data

SF-Tool can be accessed along with other programs for data validation and deposition at **deposit.pdb.org**. Questions, comments, and suggestions should be sent to **help@deposit.pdb.org**.

Many visitors came to the RCSB PDB stand for materials, demonstrations, and to talk to team members.



The ACA exhibition hall saw the debut of the new poster How Do Drugs Work? *This poster, which highlights several protein-drug PDB structures, will be distributed at the RCSB PDB booth at several upcoming meetings.*

Meeting with Depositors at the ACA

The RCSB PDB exhibited alongside the PSI Structural Genomics Knowledgebase (PSI SGKB) at the 2009 **Meeting of the American Crystallographic Association** (July 25-30; Toronto, Canada). Many data depositors and users visited the booth for the latest RCSB PDB publications, to ask questions, and to give feedback on new website features.

The scientific poster "The PDB: Updates and Future Plans" highlighted the recent improvements made to the PDB File Format and future plans for a common wwPDB deposition and annotation tool.

As part of the Crystallography Education session, a presentation on *Looking at Structures with Diverse Audiences* was given.

Deposition Statistics

In the third quarter of 2009, 2085 experimentally-determined structures were deposited to the PDB archive. The entries were processed by wwPDB teams at the RCSB PDB, PDBe, and PDBj.

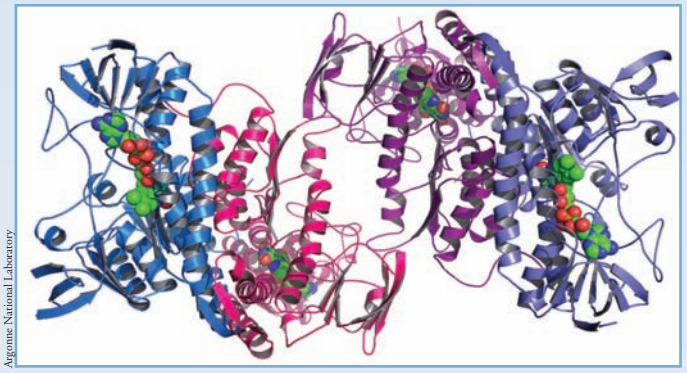
Of the structures deposited, 74.8% were deposited with a release status of HPUB; 22.6% were released as soon as annotation of the entry was complete; and 2.6% were held until a particular date. 93.0% of these entries were determined by X-ray crystallographic methods; 6.2% were determined by NMR methods.

1927 structures were released in the PDB during the same period.

Structural Genomics News

Two structural genomics centers recently made the news for reaching deposition milestones this summer.

In July, the Midwest Center for Structural Genomics (MCSG; **www.mcsg.anl.gov**) became the first Protein Structure Initiative (PSI) center to determine its 1,000th protein structure. Founded in 2000, the MCSG targets proteins from several biomedical and special classes, with a particular focus on characterization of potential novel virulence factors, pathogenic factors, and cases of viral molecular mimicry in pathogenic bacteria uncovered by bioinformatics methods as part of the NIH-funded PSI efforts.

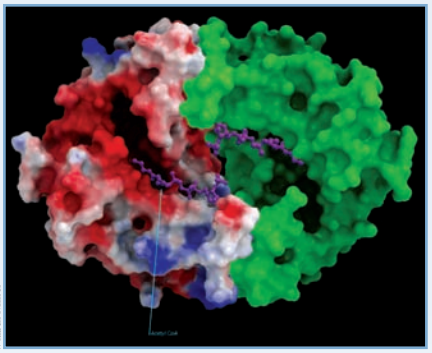


Argonne National Laboratory

The MCSG's 1,000th protein structure is a dehydrogenase from the bacterium Colwellia psychrerythraea. *The enzyme is capable of generating harmful reactive oxygen species and has been implicated in neurodegeneration, ischemia-reperfusion, cancer and several other disorders.*

PDB ID: 3ic9. *K. Tan, E. Rakowski, S. Clancy, A. Joachimiak (MCSG) The structure of dihydrolipoamide dehydrogenase from* Colwellia psychrerythraea *34H.* ***DOI:10.2210/pdb3ic9/pdb***

In August, investigators at the Center for Structural Genomics of Infectious Diseases (CSGID; **csgid.org/csgid**) determined their 100th pathogen protein structure in less than two years. Funded by the National Institute of Allergy and Infectious Diseases, the CSGID applies structural genomics approaches to potential drug targets from NIAID category A, B, and C priority pathogens, with the goal of solving 500 structures over a five-year period.



Wladek Minor

The 100th protein structure is a BA2930 protein thought to be a major part of anthrax's resistance mechanism to the aminoglycoside family of antibiotics, including streptomycin, gentamycin and kanamycin.

PDB ID: 3ijw. *M.M. Klimecka, M. Chruszcz, T. Skarina, O. Onopryienko, M. Cymborowski, A. Savchenko, A. Edwards, W. Anderson, W. Minor (CSGID). Crystal structure of BA2930 in complex with CoA.* ***DOI:10.2210/pdb3ijw/pdb***

Current structural genomics news, features, and highlights are available from the PSI Structural Genomics Knowledgebase at **kb.psi-structuralgenomics.org.**

Data Query, Reporting, and Access

Improved Navigation of the RCSB PDB Website

www.pdb.org has been reorganized to make navigating the website and search results easier and more intuitive.

The left-hand menu now groups frequently-used web pages into sections–Deposition, Home, Search, Tools, and Education–that can be moved up and down to create a left-hand menu ordered by user interest. This customized menu will then appear on every webpage. Links that appeared in the old left-hand menu system have been moved to more contextual places, such as query result pages.

By default, search results now display the most recently-released structure first. Results can also be sorted by PDB ID, Residue Count, and Resolution.

Query result pages offer different tabs for reviewing any Structure Hits, Unreleased Structures, Citations, Ligand Hits, Web Page Hits, and GO, SCOP, and CATH Hits. Options for viewing, downloading, and generating reports for these search results appear as icons at the top of each page. Mousing over the icon indicates what action will be performed. A detailed description of the new result browser options is available from the **What's New** page at www.pdb.org.

Sequence Similarity Views of PDB Structures

Interested in finding homologous protein structures, or finding a non-redundant set of proteins? The new Sequence Similarity View is an easy way to get this information based upon a protein chain in a PDB entry. This view, available through a tab at the top of each Structure Summary page, provides clusters of structures at different levels of sequence similarity. Details about the structures of all homologous proteins found within a cluster can then be examined.

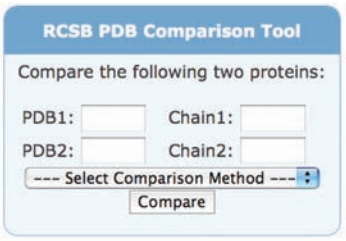
The clusters are based on a weekly BLAST analysis of all proteins with more than 20 amino acids in the PDB. Many PDB entries contain several chains, so the sequence similarity is defined on a chain-by-chain basis, with the results returned for the entire structure. For more information, see the description at www.pdb.org/pdb/statistics/clusterStatistics.do.

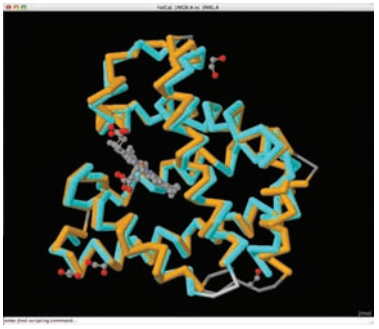
New Tool For Exploring Sequence and Structure Alignments

The new **RCSB PDB Comparison Tool** can calculate pairwise sequence and structure alignments using a variety of methods. This feature is available on the Compare Structures web page and as a downloadable web widget.

This functionality is also integrated with the Sequence Clusters offered from each entry's Sequence Similarity tab. Users can select a pair of chains from a given sequence cluster, and then run either sequence or structure alignments.

The current sequence alignments possible are: blast2seq,[2] Needleman-Wunsch,[3] and Smith-Waterman;[4] the structure alignments offered are FATCAT,[5] Mammoth,[6] TM-Align,[7] and TopMatch.[8] FATCAT alignments can be viewed on the external server at fatcat.burnham.org/fatcat or through the RCSB PDB's new JFatCat tool. JFatCat is a 3D structure alignment program based on FATCAT, BioJava[9] and Jmol.[10] JFatCat can be launched from the Comparison Tool or downloaded to the desktop.





FATCAT structure alignment for 1mgn[11] and 2nrl[12] viewed using the JFatCat tool

Beta Release of Redesigned BioSync

Synchrotron beamlines account for close to 80% of current X-ray structures deposited to the PDB. The new capabilities and descriptions of operational synchrotron beamlines worldwide are now available at biosync-beta.rutgers.edu. The new site also offers search functions for services and equipment. BioSync has been redesigned to include the dramatic changes in data collection capabilities at synchrotron beamlines (including remote data collection, mail-in, crystallization and structure solution services, robotics handling for crystal screening and mounting, microfocus beams and facilities for collecting data under extreme conditions). The new look and feel of the website helps users find information about particular beamlines and to search for capabilities, services and equipment. The website's enhanced interface lets synchrotron personnel dynamically edit and add data.

BioSync offers deposition statistics by each synchrotron site and by geographical region. Galleries of structures and tables containing citations and other general information (*e.g.*, phasing methods, resolution, R-factors, numbers of atoms) are also available. A separate set of statistical tables, galleries and informational tables is provided for structures produced by structural genomics efforts.

The redesign and upgrade of BioSync is being funded by NIGMS.

wwPDB FTP Advisory Notice

Four changes will be made to the wwPDB FTP site on November 24, 2009.

- The script for mirroring the FTP site using the rsync program will be modified to prompt the user to choose one of three rsync servers (RCSB PDB, PDBe, PDBj).
- The top-level README file will point to the download instructions hosted on the wwPDB website.
- The directory of newsletters will be updated.
- Sequence cluster data that is used only by the RCSB PDB website will be removed from the wwPDB FTP site. The data will be made available from the RCSB PDB at ftp://resources.rcsb.org/sequence/clusters/.

Detailed information can be found at www.wwpdb.org.

Outreach and Education

Poster Prizes Awarded at ACA and ISMB

ACA Poster Prize Winner Magdalena Korczynska with Victor Young and ACA President Dr. Robert Von Dreele.

Magdalena Korczynska was awarded the RCSB PDB Poster Prize for best student poster related to macromolecular crystallography at the ACA Annual Meeting for *Structural Insight into Homoserine Transacetylase from* Haemophilus influenzae (Magdalena Korczynska,[a] I. Ahmad Mirza,[a] and Albert M. Berghuis;[a,b] [a]Departments of Biochemistry and [b]Microbiology & Immunology, McGill University, Mon- treal, QC, Canada).

Many thanks to the judges: Joe Ng (The University of Alabama in Huntsville), John Rose (University of Georgia), and Emil Pai (University of Toronto).

Christoph Malisi

The award for the best student poster in the category structure function and prediction at the Joint 17th Annual International Conference for Intelligent Systems for Molecular Biology (ISMB) and the 8th European Conference on Computational Biology (ECCB; June 27-July 2; Stockholm, Sweden) went to Christoph Malisi for *Automated Scaffold Selection for Enzyme Design* (Christoph Malisi,[a,b] Oliver Kohlbacher,[b] and Birte Höcker;[a] [a]Max Planck Institute for Developmental Biology, Protein Design Group, Tübingen, Germany; [b]Division for Simulation of Biological Systems, Center for Bioinformatics, Eberhard-Karls-Universität Tübingen, Germany).

Many thanks to the judges: Marco Punta (Committee Chair, Columbia University), Yana Bromberg (Columbia University), Stefano Lise (University College London), Silvio Tosatto (Universita' di Padova, Italy), Maria Valentini (CRS4, Italy), Curtis Huttenhower (Harvard University), and Lars Arvestad (KTH Royal Institute of Technology, Sweden). Thanks also to the full reviewing committee and to the International Society for Computational Biology.

The winners will receive an educational book and a subscription to *Science* magazine.

Recent and Upcoming Meetings and Presentations

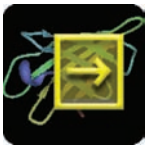
The RCSB PDB exhibited along side the PSI SG KB **at the 23rd Annual Symposium of The Protein Society** (July 25-29, Boston, MA). Lihua Tan presented the poster *A Lot More Than Coordinates: A Short Tour of a PDB File,* which highlighted all of the rich information available in a PDB file other than the X, Y, Z coordinates.

Demonstrations of new RCSB PDB features were given at the **Joint 17th Annual International Conference for Intelligent Systems for Molecular Biology (ISMB) and the 8th European Conference on Computational Biology (ECCB)** (June 27-July 2; Stockholm, Sweden). At the ISMB satellite **Bioinformatics Open Source Conference (BOSC)**, the RCSB PDB's Andreas Prlic presented *BioJava 2009: an Open-Source Framework for Bioinformatics*, and at the **BioLink Special Interest Group Session** on the Future of Scientific Publishing, Phil Bourne discussed *OpenID vs. ResearcherID.*

At the **Essentials for Educating Biochemistry and Molecular Biology Undergraduates Symposium** (August 5-8; Colorado Springs, CO) sponsored by the American Society for Biochemistry and Molecular Biology, Shuchismita Dutta ran a workshop entitled *Molecular Visualization and Protein Databases (IIA): Tools, Rules and Stories: A Protein Data Bank Workshop Series.*

The RCSB PDB will exhibit at the **American Society for Cell Biology** meeting at the end of 2009 (December 5-9; San Diego, CA) and at the **Biophysical Society** meeting at the beginning of 2010.

Turn Your Computer into a PDB Structure Kiosk



Highlight structures from your lab, institution, or class with the *Molecules in Motion Kiosk Viewer*. Using a list of PDB IDs, this full-screen animation program will display any PDB structure from different angles and perspectives. It also focuses on any chemical components within the structure. The Java viewer can be downloaded or launched from the **Educational Resources** page.

The Molecules in Motion Kiosk Viewer on display at the Busch Campus Center at Rutgers, The State University of New Jersey.

High School Teams to Build 3D Protein Models

The protein modeling event will be held at more than 20 Science Olympiad tournaments held across the country in 2010. The theme for the competitions will be influenza, with teams building models of hemagglutinin and neuraminidase.

The RCSB PDB supervises and judges the competition in New Jersey (**education.pdb.org**) and will be posting tips and news at **twitter.com/buildmodels**.

The event was developed by the Center for BioMolecular Modeling. For more information, please see **cbm.msoe.edu/stupro/so.**

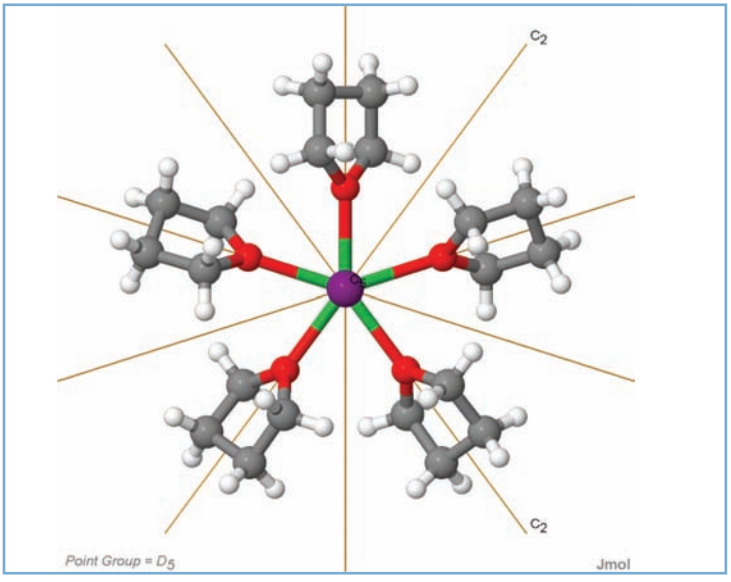
Education Corner by Gary M. Battle, Ph.D.

Symposium on the Applications of Small-molecule Crystal Structure Information in Chemical Education

As the use of crystal structure information continues to broaden, the Cambridge Structural Database (CSD) has become an indispensable resource for educators involved in both undergraduate teaching and research. This rapidly developing area was the focus of a symposium at the 238th national meeting of the American Chemical Society in Washington, D.C. on August 19, 2009. Featuring presentations from prominent educators, the symposium showcased ways in which the CSD is being used to enhance student learning across the entire span of the chemistry curriculum.

The value and availability of crystal structure information...

Frank Allen, who is now an emeritus research fellow at the CCDC, opened the symposium by highlighting the unique advantages that small molecule crystal structure information offers chemistry teachers. The benefits of visualizing and manipulating a diverse range of 'real' 3D structures onscreen and the pedagogical value of using experimentally measured data were explained. Guy Crundwell from Central Connecticut State University then discussed how using raw data mined from the CSD challenges his students to think more critically about the fundamental topics of bonding and molecular structure. In addition to this, Gary Battle from CCDC (your author) presented a number of free teaching resources, including a carefully compiled online interactive teaching subset of the CSD (**webcsd.ccdc.cam.ac.uk/teaching_database_demo.php**) and a number of associated tutorial exercises. John Woolcock of Indiana University of Pennsylvania also explained how WebCSD (the online search interface to the CSD) was used to great effect during the American Crystallographic Association (ACA) intensive ten-day summer course.



Display of symmetry elements in $YbI_2(THF)_5$ from Dean Johnston's symmetry gallery

GARY M. BATTLE *completed a Ph.D. in synthetic organic chemistry with Dr. Andrew Clark at the University of Warwick, UK in 2002, and subsequently joined the Cambridge Crystallographic Data Centre (CCDC; **www.ccdc.cam.ac.uk**) where he is now a Senior Applications and Research Scientist. In addition to providing user support and training on CCDC's full suite of software tools, Gary also leads the Centre's educational outreach activities.*

Gary's work has been focused on the Teaching Database–a free 500-structure subset of the Cambridge Structural Database (CSD). This resource offers a 500-structure teaching database of a wide variety of molecules (from adrenaline to zirconium complexes) that can be used to enhance learning across the chemistry curriculum.

*To learn about this and other teaching tools offered by the CCDC, please visit **www.ccdc.cam.ac.uk/free_services/teaching/***

Exploring concepts of symmetry...

A highlight of the symposium was a presentation from Dean Johnston of Otterbein College. Dean introduced and demonstrated the online Symmetry Resources at Otterbein College (**symmetry.otterbein.edu**). The resources contained within this excellent website are designed to help students learn concepts of molecular symmetry and to help faculty teach these concepts. A point-group symmetry tutorial with interactive displays and animations guides students through a number of symmetry elements and operations. A gallery of 70 unique molecules with interactive display of symmetry elements and animation of operations is well worth a look. Finally, the 'symmetry challenge' includes an interactive flow chart that can be used to test and follow the process of determining the point group of a particular molecule–a great way to practice. Dean used CSDSymmetry (**www.ccdc.cam.ac.uk/free_services/csdsymmetry/**), a freely available database of symmetry-related information from CCDC, to identify interesting molecules with unique point groups for inclusion on his website.

Continuing this theme, St. Olaf College's Bob Hanson demonstrated some brand-new crystallographic symmetry capabilities within Jmol (the widely used open-source Java viewer for 3D chemical structures) that were introduced specifically for this symposium. These features can be viewed in the Jmol Crystal Symmetry Explorer at **chemapps.stolaf.edu/jmol/docs/examples-11/jcse/explore.htm**

Broad appeal...

Speakers from a variety of disciplines contributed to the symposium, demonstrating the utility of crystallographic information across the whole of the chemistry curriculum.

Kraig Wheeler (Eastern Illinois University) spoke about conceptualizing reaction mechanisms using crystallographic data, and highlighted several examples of how crystallographic data can serve to support

existing reaction theories and unravel mechanistic details in the organic classroom. Katherine Kantardjieff of California State University Fullerton presented a guide for users and consumers of crystallographic information from a biochemical perspective. Stephen Koch from State University of New York at Stony Brook gave an enthusiastic talk on how students in general chemistry through to advanced inorganic chemistry classes use the CSD to explore the diverse structural chemistry of molecular inorganic compounds. Also, Virginia Pett from the College of Wooster discussed how her physical chemistry students use the CSD in computer-based laboratory sessions.

Spreading the word...

The symposium concluded with examples of how CSD-based teaching resources and materials are being disseminated and shared within the educational community.

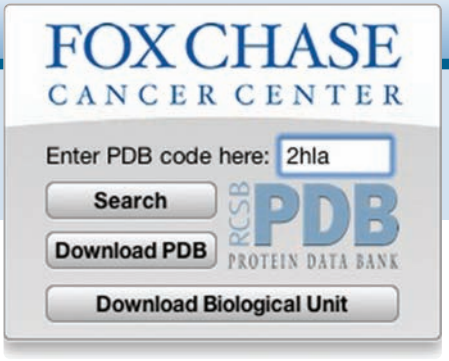
Illinois State University's Gregory Ferrence summarized a two-year project supported by an NSF-funded Discovery Corps Fellowship. The grant has provided full Cambridge Structural Database System (CSDS) access to more than 30 Primarily Undergraduate Institutions (PUIs) in the USA and has enabled Greg to visit many of these universities and colleges to provide training and workshops, and to encourage others to develop their own CSDS-based educational materials.

Barbara Reisner from James Madison University provided an overview of IONiC (Intellectual Online Network of Inorganic Chemists), a vibrant virtual 'community of practice' that facilitates collaborative development of learning materials and their dissemination to the wider inorganic community. Their website, VIPEr (Virtual Inorganic Pedagogical Electronic Resource, **www.ionicviper.org**), serves both as a repository and as a user-friendly platform that facilitates virtual collaboration and community building. VIPEr already features a number of excellent CSD-based teaching exercises and activities.

Full symposium details including copies of the presentations are available from the CCDC website **www.ccdc.cam.ac.uk/free_services/teaching**. We would also be happy to learn of any other examples

PDB Community Focus

Roland L. Dunbrack, Jr., Ph.D., Associate Professor
Program in Molecular and Translational Medicine
Fox Chase Cancer Center



Brian Weitzner and Roland Dunbrack offer an OSX Dashboard widget that provides easy downloads of PDB files and quick access to the RCSB PDB Structure Explorer web page for any given PDB ID. All of the software offered by Roland's laboratory is available at ***dunbrack.fccc.edu****.*

Q. *How do you use the PDB for your research? What do you think its value is?*

A. My group works on developing methods and software for protein structure prediction and the statistical study of protein structures required to do that well. The PDB is therefore at the center of all we do. For instance, we have used sets of thousands of structures to develop new statistical analyses of the protein backbone and of protein side chains for a new version of our backbone-dependent rotamer library. These are being used in our side-chain prediction program SCWRL4,[13] released in May 2009 and in developmental versions of the Rosetta program from David Baker's group. To train and test the accuracy of our structure prediction methods, we also require large sets of protein structures from the PDB.

Of course, large-scale study of protein structures, sometimes known as structural bioinformatics, has a lot of purposes, and there are many research groups worldwide who do such work. So a few years ago we developed the PISCES server,[14-15] which allows users to produce lists of PDB entries or chains from PDB entries with user-selected criteria such as resolution, R-factors, and maximum sequence identity between any two proteins in the output list. A very useful feature is that the user can input his or her own list of PDB chains, say, all kinases or all PDB entries with available structure factors, and get a culled list from those with the desired resolution and sequence identity cutoffs. We have used this feature ourselves many times.

Q. *The Fox Chase Cancer Center houses a variety of scientists and physicians. How does your research group interact with all of them?*

A. It's part of the culture and history of my institution that we are a collaborative and congenial group of faculty. When I first got to Fox Chase, one of my colleagues showed me some functional results on mutants of an enzyme he worked on. These mutants were associated with homocystinuria in people, which can be a dangerous condition. Some patients were responsive to vitamin B6, the enzyme's cofactor, and others were not. So we made a model of the protein based on a 19% sequence-identical homologue and the most unresponsive mutations were located in the active site, while the others were mostly on the surface. He was astonished that I could tell him this information, in the absence of a crystal structure of his protein. He gave one of our weekly faculty seminars and showed the model, and not long after I had people knocking on my door. We had some really nice collaborations in the next few years, and in 2003 the Center was awarded a grant from the Pew Charitable Trusts in Philadelphia to establish the Fox Chase Molecular Modeling Facility with one full-time staff member. That person's job is to use many kinds of modeling methods that are publicly available to model proteins and protein complexes to help explain existing experimental data and to generate new hypotheses and experiments. We interact frequently with the investigators during each project so that they can make the most use of the models and the analysis, and we provide text and images for papers and grant proposals.

In the 12 years I have been at Fox Chase, my group has worked with about 60 of my colleagues on over 120 different protein targets. Questions range from *where are the functional domains in this 2000 amino acid cancer-associated protein?* to *where can I make a mutation to knock out one protein-protein interaction of my protein without affecting others?* Adrian Cantuescu in the Facility developed a graphical user interface, MolIDE, to make basic homology modeling (searching the PDB, selecting a template, producing and editing the sequence-template alignment, and loop and side-chain modeling) very easy.[16] We are now extending it to complexes of proteins.

Q. ***In addition to your work in protein sequence analysis and structure prediction, you have been focused on the accuracy and representation of biological assemblies. Why do you think this is so important? How did this interest develop?***

A. This interest developed largely from working with my colleagues. We had many cases where we needed to build complexes from dimers to octamers, and while we could do that with some manual intervention, it was sometimes a tedious process. We also had some cases where deleterious mutants were obviously in protein-protein interfaces, and we wanted it to be easier to make such models. Initially, I naively thought the available databases, the PDB itself and the Protein Quaternary Server at the EBI would have very similar sets of biological units for PDB entries. It turned out they agreed with each other only about 80% of the time. So we produced a database-software program, ProtBUD, to be able to retrieve and present in a sortable table all the biological unit information from both sources for any query protein family.[17] The program also provides ligand information, making it easy to find perhaps the single entry among dozens in a protein family that has the biological ligand of interest, instead of having to search all the entries manually.

We also were working on a model of a sulfotransferase, SULT4A1, for one of my colleagues and she showed us a paper on crosslinking, protease digestion, and mass spectrometry of a dimer of this protein locating residues involved in the homodimer interface. It turned out the interface was present in the two crystal structures then available for SULT family members when the paper was published (2001). In 2005, there were something like 12 different crystal forms of various family members, and we found the same dimer interface in every single one of them.[18] Only one of 20 or so PDB entries had the dimer interface annotated in the PDB. The rest were either monomeric or had a bunch of other different interfaces, not shared by any more than one crystal form. PQS had one or two interfaces correct, but only when the dimer was in the asymmetric unit. This led us to do a PDB-wide examination of crystals in protein families, to identify common interfaces.[19] Annotations in the PDB at that time were mostly from the authors or PQS, and we found that for families with a large number of crystal forms containing the same interface, about 90% of the entries had that interface in the biological assembly in the PDB. For the newly developed PISA program at the EBI, the number was about 95%. The PDB is now including the PISA annotations in its biological assemblies, which I think is a great idea. It gives users an opportunity to examine various hypothetical assemblies for any entry they may be interested in.

Q. ***DashPDB is a recent addition to a variety of programs and resources available from your website for download. Recently, you released an updated version of SCWRL, a program for predicting protein side-chain conformations. How do you think your software programs are being used?***

A. I strongly believe in making the methods and programs developed by computational biology research groups available to the public either as webservers or downloadable software or depending on the purpose, preferably both. There are many papers on method development with no software available, or only a webserver so that a user can only do one manually input request at a time, which limits the kind of studies that can be done. Sometimes software is available, but has to be compiled by the user or is written such that the program is very finicky about complicated input files in ways that are not well documented. So we try to provide software that is easy to use and does what it says it does with good documentation. It's easier to do this if we think about this at the beginning of method development, rather than as an afterthought when a student or postdoc has created something that nobody else can use.

Our software gets used in many different ways. SCWRL is easy enough to use that I think groups interested in things other than just structure prediction use it. We have about 3300 licensees. MolIDE, which makes models with SCWRL from a query sequence, has about 1200 users. Since SCWRL is very fast, it gets used on some servers that provide sequence-template alignments for remote homologues (like FFAS) in order to make a quick model based on the alignment. I also hope our programs get used in the kinds of productive and very enjoyable collaborations we have been fortunate enough to have over the last 12 years.

References

1. A.A. Vaguine, J. Richelle, S.J. Wodak. (1999) SFCHECK: A unified set of procedures for evaluating the quality of macromolecular structure-factor data and their agreement with the atomic model. *Acta Crystallogr* **D55**:191-205.
2. T.A. Tatusova & T.L. Madden. (1999) BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol Lett* **174**:247-250.
3. S.B. Needleman & C.D. Wunsch. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48**:443-453.
4. T. F. Smith & M. S. Waterman. (1981) Identification of common molecular subsequences. *J Mol Biol* **147**:195-197.
5. Y. Ye & A. Godzik. (2003) Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics* **19**:ii246-255.
6. A.R. Ortiz, C.E. Strauss & O. Olmea. (2002) MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci* **11**:2606-2621.
7. Y. Zhang & J. Skolnick. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res* **33**:2302-2309.
8. M. J. Sippl & M. Wiederstein. (2008) A note on difficult structure alignment problems. *Bioinformatics* **24**:426-427.
9. R.C. Holland, T.A. Down, M. Pocock, A. Prlic, D. Huen, K. James, S. Foisy, A. Drager, A. Yates, M. Heuer & M.J. Schreiber. (2008) BioJava: an open-source framework for bioinformatics. *Bioinformatics* **24**:2096-2097.
10. Jmol: an open-source Java viewer for chemical structures in 3D. **http://www.jmol.org/**
11. M.S. Hargrove, E.W. Singleton, M.L. Quillin, L.A. Ortiz, G.N. Phillips Jr., J.S. Olson, A.J. Mathews. (1994) Tyr apomyoglobin as a reagent for measuring rates of hemin dissociation. *J Biol Chem* **269**:4207-4214.
12. E.R. Schreiter, M.M. Rodriguez, A. Weichsel, W.R. Montfort, J. Bonaventura. (2007) S-nitrosylation-induced conformational change in blackfin tuna myoglobin. *J Biol Chem* **282**: 19773-19780.
13. G. G. Krivov, M. V. Shapovalov & R. L. Dunbrack, Jr. (2009) Improved prediction of protein side-chain conformations with SCWRL4. *Proteins* 77:778-795.
14. G. Wang & R. L. Dunbrack, Jr. (2003) PISCES: a protein sequence culling server. *Bioinformatics* **19**:1589-1591.
15. G. Wang & R. L. Dunbrack, Jr. (2005) PISCES: recent improvements to a PDB sequence culling server. *Nucleic Acids Res* **33**:W94-98.
16. A. A. Canutescu & R. L. Dunbrack, Jr. (2005) MollDE: a homology modeling framework you can click with. *Bioinformatics* **21**:2914-2916.
17. Q. Xu, A. Canutescu, Z. Obradovic & R. L. Dunbrack, Jr. (2006) ProtBuD: a database of biological unit structures of protein families and superfamilies. *Bioinformatics* **22**:2876-2882.
18. B. Weitzner, T. Meehan, Q. Xu & R. L. Dunbrack, Jr. (2008) An unusually small dimer interface is observed in all available crystal structures of cytosolic sulfotransferases. *Proteins: Structure, Function and Genetics* **75**:289-295.
19. Q. Xu, A. A. Canutescu, G. Wang, M. Shapovalov, Z. Obradovic & R. L. Dunbrack, Jr. (2008) Statistical analysis of interface similarity in crystals of homologous proteins. *J Mol Biol* **381**:487-507.

RCSB PDB Partners

The RCSB PDB is managed by two partner sites of the Research Collaboratory for Structural Bioinformatics:

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

The RCSB PDB is a member of the Worldwide Protein Data Bank (www.wwpdb.org)

RCSB PDB Management

DR. HELEN M. BERMAN, Director
Rutgers, The State University of New Jersey
berman@rcsb.rutgers.edu

DR. MARTHA QUESADA, Deputy Director
Rutgers, The State University of New Jersey
mquesada@rcsb.rutgers.edu

DR. PHILIP E. BOURNE, Associate Director
San Diego Supercomputer Center and the Skaggs School of Pharmacy and Pharmaceutical Sciences,
University of California, San Diego
bourne@sdsc.edu

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

STATEMENT OF SUPPORT: *The RCSB PDB is supported by funds from the National Science Foundation, the National Institute of General Medical Sciences, the Office of Science, Department of Energy, the National Library of Medicine, the National Cancer Institute, the National Institute of Neurological Disorders and Stroke, and the National Institute of Diabetes & Digestive & Kidney Diseases.*

RCSB PROTEIN DATA BANK
www.pdb.org • info@rcsb.org
Department of Chemistry and Chemical Biology
Rutgers, The State University of New Jersey
610 Taylor Road
Piscataway, NJ 08854-8087
USA

Return Service Requested