Fall 2007 • Number 35

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

RCSB PDB PROTEIN DATA BANK NEWSLETTER

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

Contents

SNAPSHOT: OCTOBER 1, 2007

46051 released atomic coordinate entries

MOLECULE TYPE		EXPERIMENTAL TECHNIQUE	
42350	proteins, peptides, and viruses	39184	X-ray
1787	nucleic acids	6621	NMR
1881	protein/nucleic acid complexes	154	electron microscopy
33	other	92	other
		28451	structure factor files
		3648	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)

Message from the RCSB PDB

In September, the RCSB PDB hosted a series of important meetings at the Chauncey Conference Center in Princeton, New Jersey. These meetings, held back-to-back, provided an unprecedented opportunity for discussion and planning.

The RCSB PDB Advisory Committee, an organization of international experts in X-ray crystallography, NMR, 3-D EM, bioinformatics, and education, listened to presentations and discussed future plans and goals. This review was followed by the Worldwide Protein Data Bank's Advisory Committee (wwPDB AC) Meeting. This panel of expert structural biologists includes representatives from the International Union of Crystallography and the International Conferences on Magnetic Resonance in Biological Systems. After the advisory committee considered reports from the wwPDB, a "Funding Forum" took place. In this session, the wwPDB AC sought advice from the representatives present from the agencies that fund the individual groups about funding options for the continued operation of the wwPDB organization. Representatives from the academic and industrial research communities that rely on the PDB for their research efforts also described of the PDB value for those present.

To take advantage of having so many people in the same place, a retreat was held for members from all of the wwPDB sites. September's retreat was attended by nearly 50 people from the four groups. While the wwPDB sites interact regularly, this was the first meeting on such a large scale. Many colleagues met in person for the first time. The retreat also provided an opportunity to celebrate the August release of the remediated PDB archive (ftp://ftp.wwpdb.org). This wwPDB milestone represents years of work and unprecedented international collaboration.

At the start of the meeting, the wwPDB team was treated to presentations from PDB users and advisors–Edward N. Baker (Professor of Structural Biology, University of Auckland), Angela Gronenborn (UPMC Rosalind Franklin Professor and Chair, Department of Structural Biology, University of Pittsburgh), Gerard Kleywegt (Research Fellow of the Royal Swedish Academy of Sciences, Research Scientist, Uppsala University), Marin Van Heel (Professor of Structural Biology, Imperial College London), and Soichi Wakatsuki (Professor, Structural Biology Research Center, High Energy Accelerator Research Organization, Japan). The retreat then focused on discussing how the wwPDB could evolve with and anticipate the needs of the scientific community.

Special thanks to all of the advisors, funding agency representatives, and wwPDB collaborators who traveled to New Jersey for these important meetings.

Members from the RCSB PDB, MSD-EBI, PDBj, and BMRB at the wwPDB Retreat

Data Deposition and Processing

Structure Deposition Overview

Structures can be deposited to the wwPDB using the tools ADIT, ADIT-NMR, or AutoDep.

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.

The following was presented this summer by Lead Annotator Jasmine Young at the American Crystallographic Association's Annual Meeting.

5 Easy Steps for Fast, Accurate, and Complete Data Deposition using the ADIT system

The first step would be to **Verify the Sequence (Step 1)** you are depositing. The sequence submitted to the PDB archive should include any residues missing due to lack of electron density, cloning artifacts and HIS tags that were not cleaved, and any mutations or substitutions. This protein and or nucleotide sequence should be entered into a sequence database (*e.g.*, BLAST)[1] to be compared with any existing sequence database references. When depositing, you should check sequence database correspondences and fix any sequence discrepancies found, such as unobserved gaps.

This sequence should then be entered, along with other information, when you **use the pdb_extract tool (Step 2)**[2]. This program takes data from crystal or NMR structure determination programs and automatically fills in items necessary for deposition–refinement statistics, data collection, phasing, and more–to generate a complete data file for deposition. pdb_extract can also be used to build a text file that can be edited to use when depositing many related structures.

The mmCIF file generated from pdb_extract and for crystal structures, the structure factor file, should be uploaded into the **Validation Server (Step 3)**[3]. (For depositors not using pdb_extract, the Validation Server can also read PDB or mmCIF files from many refinement programs). The Precheck step looks to see if the coordinate file and structure factor file is in a readable format. The Validation step produces a report that contains information about any close contacts; bond distance and angle deviations; chirality errors; sequence/coordinate (mis)alignments; missing and extra atoms or residues; and distant waters. The report also includes output from the programs NUCheck, SFCHECK[4], and MolProbity[5]. Depositors should carefully review these reports and make corrections where necessary.

When depositing structures with ligands, users should **Check Any Ligands (Step 4)** to find chemical component IDs (three letter codes) for existing ligands. If the ligand is not found in the Chemical Component Dictionary, a chemical diagram (with bond order), IUPAC name, synonyms, formula, and potential three letter code should be emailed to **deposit@deposit.rcsb.org**

Once you are comfortable with the reports generated from the Validation Server and you have either found or submitted your ligand, the structure can be deposited using **ADIT (Step 5)**.[6]

The web-based version of ADIT asks users to upload the coordinate file and experimental data. The tool can then be used to enter any missing information. At this point, you should also indicate if the structure's sequence can be released before the entire entry is released. After the deposition has been reviewed, the structure can be submitted. A PDB ID will be returned automatically.

A desktop version of ADIT is available for users who are behind firewalls.

After your entry is deposited, the annotation staff will work to represent your PDB data in the best possible way. This process involves:

- Reviewing the entry for self-consistency
- Confirming that the entry title matches the content of the deposited entry
- Correcting any format errors in data and coordinates
- Checking the sequence
- Inserting sequence database references
- Providing a protein name and synonyms
- Checking the scientific name of the source organism
- Confirming the chemical consistency between ligand name and coordinates
- Adding biological assembly information
- Visually checking the structure
- Generating and reviewing validation reports
- Finding citation references with PubMed[7]

Any findings are compiled and sent back to the depositor. If no problems are found with the entry by the annotator or the depositor, then it is considered automatically approved and is ready to be released based upon the deposited release status. Entries can be released immediately, held until publication of the corresponding primary citation, or held until a particular date. Depositions cannot be held longer than one year.

Entries currently released by the wwPDB follows the PDB format as described in the PDB Contents Guide Version 3.1 and the mmCIF format that complies with the current PDB Exchange Dictionary (PDBx) v1.045. These formats contain the new features incorporated as part of the Remediation Project, including:

- Two letter codes for DNA labeling
- Better representations for complex assemblies
- Standardized atom nomenclature that follows IUPAC naming

Deposition updates and questions about this process should be sent to **deposit@deposit.pdb.org**.

Deposition Resources

1. **BLAST**
   www.ebi.ac.uk/blast2 or
   www.ncbi.nih.gov/BLAST
2. **pdb_extract**
   pdb-extract.rcsb.org
3. **Validation Suite**
   deposit.rcsb.org/validate or
   pdbdep.protein.osaka-u.ac.jp/validate
4. **Chemical Component Dictionary**
   remediation.wwpdb.org/downloads.html
5. **ADIT**
   deposit.rcsb.org/adit or
   pdbdep.protein.osaka-u.ac.jp/adit

   ADIT-NMR
   deposit.bmrb.wisc.edu/bmrb-adit or
   nmradit.protein.osaka-u.ac.jp/bmrb-adit

1. S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**:403-410.
2. H. Yang, V. Guranovic, S. Dutta, Z. Feng, H.M. Berman, and 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.
3. J. Westbrook, Z. Feng, K. Burkhardt, and H.M. Berman (2003) Validation of protein structures for the Protein Data Bank. *Meth Enz.* **374**:370-385.
4. A.A. Vaguine, J. Richelle, and 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.
5. I.W. Davis, L.W. Murray, J.S. Richardson, and 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.
6. S. Dutta, K. Burkhardt, W.F. Bluhm, and H.M. Berman (2005) Using the tools and resources of the RCSB Protein Data Bank. *Current Protocols in Bioinformatics:* **1.9.1-1.9.40**.
7. 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, and E. Yaschenko (2007) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **35**(Database issue):D5-12.

Annotation Job Opening at the RCSB PDB

In addition to curating data, annotation staff at the RCSB PDB are involved in a variety of educational and outreach projects, attend professional society meetings, and assist in software development. This position offers the opportunity to participate in an exciting project with significant impact on the scientific community.

To apply, please send your resume to Dr. Helen M. Berman at pdbjobs@rcsb.rutgers.edu.

wwPDB Annotators at the September Retreat

Data Query, Reporting, and Access

New Query and Reporting Capabilities and Features

Since the RCSB PDB website and database utilize data from the wwPDB Remediation Project, queries now return more accurate results. New developments in query and reporting features also provide improved access to these data.

• Access to Remediation and Pre-remediation Data

All data in the PDB archive (ftp://ftp.wwpdb.org) reflect the new features incorporated as part of the wwPDB Remediation Project, including standardized IUPAC nomenclature[1] for chemical components. These data have been incorporated into the RCSB PDB website and database to provide improved searching and reporting capabilities. Access to the unremediated data is possible for individual structures and for the entire archive.

The left hand menu of each Structure Summary page provides download options for either remediated or unremediated data in a variety of formats. The Remediation Tab will appear on this page to describe any changes to chain and residue naming conventions made for consistency in the archive. An example description would be *This structure's single unnamed chain was assigned chain id A.*

A snapshot of the entire unremediated PDB archive (as of July 31, 2007) is available at ftp://ftp.rcsb.org. This archive will not be updated.

• 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 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.

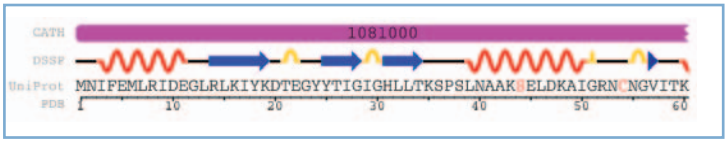
A broad range of subqueries is available including sequence searches; Gene Ontology (GO)[2] assignments; SCOP[3] and CATH[4] domain assignments; and author name searches.

These subqueries may be combined into a complex query by searching "all" or "any" of the user-specified subqueries.

• Improved Sequence Details

The Sequence Details tab offers a customizable report that displays polymer chain sequences annotated with properties such as domain and secondary structure. This feature utilizes data from the Remediation Project to provide an exact mapping of the structure sequence to the UniProt[5] sequence. Annotations from CATH, DSSP,[6] PDP,[7] and the author-approved secondary structure can be applied to either the sequence in UniProt or in the PDB entry's SEQRES information.

The size of the report can be customized for use in presentations.



The first 60 residues of T4 lysozyme (PDB ID: 108l)[8] are mapped to their entries in CATH, DSSP, and UniProt. Mutations in the sequence are shown in red.

These tabs offer different ways of exploring search results.

• Search Result Tabs

Keyword or Advanced Searches will also return different ways of exploring the search results list. Options available from the tabs shown above the default results list include:

– Citations: 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 provides a PubMed-like list of the primary citations for the structures that match a query.

– Ligand Hits: This tab lists the ligands known to interact with the structure matching the query. For example, a keyword search for "protein kinase" will return all ligands known to bind protein kinases. Linked images, names, IDs, and formulas appear for each ligand.

– Web Page Hits: Any of the more than 900 curated web pages found at the RCSB PDB website, including *Molecule of the Month* features, that contain a requested keyword are found on this tab.

– GO, SCOP, CATH Hits: These tabs link to the structures that have the same mapping in the GO, SCOP, and CATH resources. Entries are returned in a tree browser that indicates where these structures reside in the respective hierarchies. The SCOP tab, for example, indicates which hits belong to which class of proteins.

1. J.L. Markley, A. Bax, Y. Arata, C.W. Hilbers, R. Kaptein, B.D. Sykes, P.E. Wright, and K. Wüthrich (1998) Recommendations for the presentation of NMR structures of proteins and nucleic acids. IUPAC-IUBMB-IUPAB Inter-Union Task Group on the standardization of data bases of protein and nucleic acid structures determined by NMR spectroscopy. *Pure & Appl. Chem.* **70**:117-142.
2. The Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. *Nature Genetics* **25**:25-29.
3. 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**(1):257-259.
4. 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.
5. The UniProt Consortium (2007) The Universal Protein Resource (UniProt). *Nucleic Acids Res.* **35**(Database issue):D193-7.
6. W. Kabsch and C. Sander (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers.* **22**:2577-2637.
7. N. Alexandrov and I. Shindyalov (2003) PDP: protein domain parser. *Bioinformatics.* **19**(3): 429-30.
8. M. Blaber, X.J. Zhang, B.W. Matthews (1993) Structural basis of amino acid alpha helix propensity. *Science* **260**:1637-1640

Website Statistics

Access statistics for **www.pdb.org** for the third quarter of 2007

MONTH	UNIQUE VISITORS	NUMBER OF VISITS	BANDWIDTH
JUL 07	93,719	244,152	592.23 GB
AUG 07	87,494	225,482	380.69 GB
SEP 07	118,631	294,060	482.76 GB

Outreach and Education

Meeting Report: ACA, ISMB, BSR, and ACS

• Thanks to everyone who stopped by the RCSB PDB exhibit booth for demonstrations of the RCSB PDB website and discussions about the remediated data at the American Crystallographic Association's Annual Meeting (ACA; July 21-26 in Salt Lake City, UT). We also appreciate those who viewed the poster *Remediation of the PDB Archive*.

The session *Informatics in Structural Biology*, organized by John Westbrook (RCSB PDB) and Kim Henrick (MSD-EBI), focused on the applications of structural informatics and inspired a lot of interesting conversations.

Annotator Jasmine Young's presentation at the *Fun Lectures for Young Scientists* symposium is described in this newsletter (page 2).

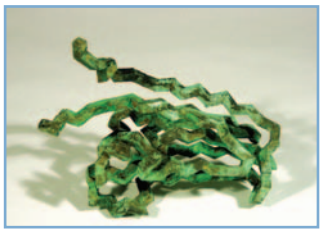
The Art of Science *exhibit was part of the ISMB meeting (July 19-25; Vienna, Austria). Special thanks to Steven Leard, BJ Morrison, and Burkhard Rost for their support and help with the exhibit and the poster prize.*

• Demonstrations of the RCSB PDB website and the *Art of Science* exhibit were found at the 15th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB) & 6th European Conference on Computational Biology.

• A poster about the BioSync resource (**biosync.rcsb.org**) was presented at the 9th International Conference on Biology and Synchrotron Radiation (BSR) by Judith L. Flippen-Anderson (August 13-17; Manchester, UK).

• At the American Chemical Society's National Meeting (ACS), Shuchismita Dutta presented a poster describing the information contained in the remediated Chemical Component Dictionary and how this dictionary was used by the wwPDB to help remediate the PDB archive (August 19-23; Boston, MA).

Art of Science Update



The Art of Science *also hosts works by the protein sculptor Julian Voss-Andreae, including this sculpture of a piece of a virus shell. For more information about Julian, please read the Winter 2007* RCSB PDB Newsletter's *Community Focus. Shown is* Virus Capsid *(2003, Cast and fabricated bronze, length 9").*

The *Art of Science* is a traveling exhibit of images from the RCSB PDB website and the *Molecule of the Month*.

This exhibit was recently hosted by the International Society for Computational Biology at the ISMB meeting. The show also was on display from September 25 – October 5 at The University of Texas Southwestern Medical Center in Dallas, Texas. That show was sponsored by the Molecular Biophysics Graduate Program, and presented in conjunction with the Molecular Biophysics' *Meet the Program* Poster Presentations.

If you would be interested in sponsoring this exhibit at your institution, please let us know at **info@rcsb.org**.

2007 RCSB PDB Poster Prizes Awarded at ACA, ISMB, and ECM

Thanks to everyone who participated in the recent RCSB PDB Poster Prize competitions. The creators of the best student posters related to macromolecular crystallography the ACA's Annual Meeting and the European Crystallographic Association's Meeting (ECM) and in the *Structure and Function Prediction* category at the ISMB meeting were awarded a subscription to *Science* and a related book.

The same prize will also be awarded at the Asian Crystallographic Association meeting later this year.

ISMB (July 19-25; Vienna, Austria)
Keren Lasker for *Determining the configuration of macromolecular assembly components based on cryoEM density fitting and pairwise geometric complementarity* (Keren Lasker, Tel Aviv University and University of California, San Francisco; Maya Topf, Birkbeck College, University of London; Andrej Sali, University of California, San Francisco; Haim Wolfson, Tel Aviv University)

Judges: Yanay Ofran (Columbia University), Predrag Radivojac (Indiana University), Alejandro Giorgetti (University of Rome), Riccardo Percudani (Universita di Parma), Michael Tress (Spanish National Cancer Research Centre), and Sean Mooney (Indiana University School of Medicine)

Poster Prize Chairman: Marco Punta (University of Georgia)

ACA (July 21-26; Salt Lake City, UT)
Hasan Demirci for *Structure Based Protein Enginering of Ribosomal Protein Trimethyltransferase* (Hasan Demirci, Steven T. Gregory, Albert E. Dahlberg, Gerwald Jogl. Department of Molecular Biology, Cell Biology, and Biochemistry, Brown University)

Judges: Mitchell J. Guss (University of Sydney), Peter Horanyi (University of Virginia), Thomas Koetzle (Argonne National Laboratory), James Phillips (Duke University Medical Center), Bernard Santarsiero (University of Illinois at Chicago), and Timothy Umland (Hauptman-Woodward Medical Research Institute)

Poster Prize Chairman: John Rose (University of Georgia)

ECM (August 22-27; Marrakech, Morroco)
Humberto Couto Fernandes for *Yellow lupine pathogenesis-related protein as a reservoir for cytokinins* (Humberto Fernandes, Anna Bujacz, Oliwia Pasternak, Grzegorz Bujacz, Michal Sikorski, Mariusz Jaskólski, Center for Biocrystallographic Research, Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland)

Judges: Alexander Wlodawer (National Cancer Institute at Frederick), Wolfram Saenger (Freie Universität Berlin), Vilmos Fulop (University of Warwick), Tomitake Tsukihara (Osaka University)

Poster Prize Chairman: Anders Liljas (Lund University)

Special thanks to John Helliwell and Petra Bombicz for their help with organizing the prize at this meeting.

EDUCATION CORNER by Dr. Melissa Kosinski-Collins, Brandeis University

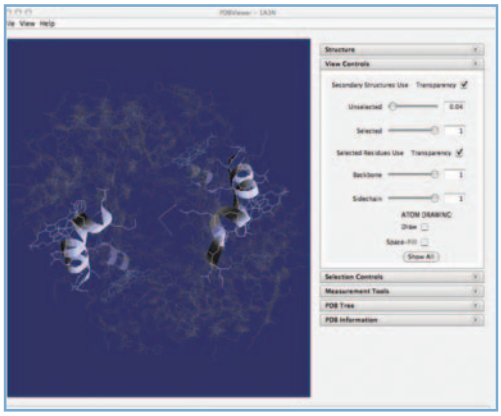
Navigating the Molecular Universe in 3D: Teaching biology students protein structure-function relationships using StarBiochem

Why does a protein do its job in the cell? Because of its shape and chemistry. As biologists, we understand and appreciate the overwhelming knowledge value in these simple statements, but as educators, we battle to try to make the protein structure-function relationship clear to our students. It is easy for us to understand this now, but to get through to our students we need to remember back to a time when we did not understand. How did we originally learn this? The answer is simple: practice.

It is clear that a very efficient way to teach to the structure-function relationship comes from letting students view some of the many deposited PDB molecules in a 3D environment. Many of the stereotypically structural "ah-ha" moments come from this type of hands-on interaction with the molecule. Students can not only identify binding pockets and partners, see disease-associated mutations, and observe structural contexts, but they can physically manipulate and, in a sense, *control* the molecule in real-time. Students of introductory biology need these types of hands-on experiences as well as practice with multiple molecules to really "get" structural biology.

Implementing such an interactive yet understandable series of exercises in the average college-level introductory biology course is a daunting task for many reasons. These hurdles include class size, computer and technology access both in the classroom and at home, time devoted to the topic in the syllabus, time involved in creating this type of homework, and the level of understanding of the incoming student. Although there are many freely available software packages that allow the students to explore in 3D, few present the material in a format that makes sense to the average biology student and are simple enough so that the student can use the program outside of the classroom on their own for additional practice.

In 2004, a project was begun at MIT to create a new program that filled the pedagogical void left in the world of structural biology. We wanted to create a viewer and a series of exercises that presented structures and functions in the same way we presented them in class that was usable outside of the classroom without staff supervision, and that allowed students many of the freedoms and exploratory options of the research-level PDB viewers. This beta version of this software was named StarBiochem.



Hemoglobin as viewed using StarBiochem (1a3n: J.R. Tame, B. Vallone (2000) The structures of deoxy human hemoglobin and the mutant Hb Tyrα42His at 120 K. Acta Crystallogr.,Sect.D ***56****:805-811)*

StarBiochem has one particular option that has become paramount to its success in the context of biology education. In class we invariably introduce protein structure as a build-up of primary, to secondary, to tertiary, to quaternary structures. Most software packages avoid mention of these levels altogether leaving the student to wonder where the levels fit in and how they are related to 3D structure they see on the screen. StarBiochem can open any protein PDB coordinate file and categorize it into these different levels allowing the student to conceptually analyze the 3D structure that they see on their screen. In the examples of hemoglobin and sickle cell anemia, the student is asked to look first at the primary structure change in the molecule, but then to determine at which structural level the disease manifests itself. Using the program as a conceptual guide, the students are asked to understand that the primary structure change from glutamic acid to valine at position 6 does not manifest as a disease until you see a change of intermolecular interaction chemistry in the quaternary structure.

StarBiochem was first piloted in an HHMI-sponsored high school field trip at MIT in March 2006. A series of guided exercises led students through an in-depth exploration of proteins with defined structure-function relationships, like sucrose-specific porin and hemoglobin. For example, the students were asked to look at the barrel-like structure of porin and reflect on how that shape might be conducive for molecular transport. The students were further asked to investigate the outer chemistry of the molecule and determine how the hydrophobic exterior of the protein might influence the ability of the protein to remain stable in its cellular membrane-bound location. StarBiochem was found to be an effective, and easy-to-use teaching tool in this context and is now being used by several of the visiting teachers as a curricular tool in their classroom. The StarBiochem high school initiative is now being further disseminated in MIT and Harvard's Broad Institute Outreach Program.

STARBIOCHEM *is the result of collaboration between the Department of Biology, the Academic Computing Group of the Department of Information Services and Technology, and the Department of Physics at the Massachusetts Institute of Technology (MIT). The founding members of the project include Dr. Melissa Kosinski-Collins, Dr. Graham Walker, Dr. John Belcher, Michael Danziger, Charles Shubert, and Ivica Ceraj. We have been privileged to be assisted by many other talented individuals including Andrew McKinney, Justin Riley, Violeta Ivanova, Professor Dan Hastings (Dean of Undergraduate Education at MIT), Dr. Vijay Kumar (Associate Dean and Director of the Office of Educational Innovation and Technology), and Dr. Jerry Grochow (MIT Vice President of Information Services and Technology). We further have been supported by the educational efforts of Dr. Julia Khodor, Dr. Megan Rokop, Dr. Mandana Sassafar, and Dr. Robyn Tanny in introductory biology courses at MIT and in the high school outreach efforts.*

Funding for this work was provided in part by the Department of Information Services and Technology at MIT, a Howard Hughes Medical Institute (HHMI) Professorship Grant awarded to Graham Walker, and a grant from the Davis Educational Foundation Grant awarded to John Belcher. The Academic Computing group that participated is this project is now the Software Tools for Academics and Researchers (STAR) group in Office of Educational Innovation and Technology for the Dean of Undergraduate Education at MIT.

MELISSA KOSINSKI-COLLINS *is now an Assistant Professor of Biology at Brandeis University.*

StarBiochem has become an integral part of the introductory biology series at MIT. We have been successful using this program on problem sets and in practice problem assignments. Most of the exercises ask students to visualize a protein that has either been discussed in class or that has a disease connection in StarBiochem. The students explore the 3-dimensional structure of the protein, the stabilizing interactions, and are then asked to relate this information to the biological function of the molecule. The students are given both graded problems and practice problems and have, thus far, been able to use this software on their own at home or in their dorm rooms. We have now moved into the exploration of nucleic acid structures as well. We have developed problems that teach the students how transcription factors and polymerase bind to and function with nucleic acids all based on shape and chemistry.

More recently, we have begun to push the boundaries of undergraduate understanding of the protein structure-function relationship even further. Starting in fall of 2007, all introductory biology students at Brandeis University will be asked to use StarBiochem to explore the structure of human eye lens protein. They will analyze the shape and chemistry of the molecule in 3D and will then select an amino acid they feel is important to the folding and structure of this protein based solely on their 3D investigation. They will be asked to think about the size and shape of the amino acid, as well as the importance and strength of the interactions in which the amino acid participates. Using site-directed mutagenesis, the students will create their own engineered protein with the ultimate goal of finding and analyzing how or if that residue actually disrupts the structure and the function of the molecule.

With the availability of so many structures, let alone so many different molecules in the PDB, it is our duty as educators to make sure that even the most inexperienced of biology students get opportunities to "see" the 3D molecular world. We have found both at Brandeis and at MIT that StarBiochem gives students the opportunity to explore the structure-function relationship, and challenges them to go one step further as well. StarBiochem engages students, reinforces classroom concepts, and encourages them to learn as research scientists do; by practice and exploration.

StarBiochem is freely available for download at **web.mit.edu/star/biochem**.

Questions about StarBiochem may be sent to **kosinski@brandeis.edu**.

PDB Community Focus:

Dr. Philip E. Bourne, RCSB PDB

PHILIP E. BOURNE *is a Professor in the Department of Pharmacology at the University of California, San Diego, co-director of the RCSB PDB, and an Adjunct Professor at the Burnham Institute and the Keck Graduate Institute.*

He received his Ph.D. in chemistry from the Flinders University of South Australia in 1980 where he studied the structural and electrophilic effects of substitution on fully saturated caged hydrocarbon molecules. While a post-doctoral fellow at Sheffield University UK he contributed to the understanding of the structural role of ferritin in iron storage. Later as a Senior Research Scientist at Columbia University, he proposed mechanisms for the role of caracurines and snake toxins that operate postsynaptically. During the 80's, first as the Director of the Cancer Center Computer Facility, and later Director of the Medical School Computer Facility at Columbia, he helped establish a tumor registry and various applications and databases in support of patient care. As a Senior Associate of the Howard Hughes Medical Institute in the early 90s, he worked on developing high performance hardware and software for computational structural biology. He moved to UCSD in 1995 to work on structural bioinformatics. His current research interests are in structural genomics, the structural basis of evolution and immunology, apoptosis, cell signaling, data and knowledge modeling and scientific visualization.

Bourne is an elected Fellow of the American Medical Informatics Association and past President of the International Society for Computational Biology. He is the Founding Editor-in-Chief of the open access journal PLoS Computational Biology, *on the Advisory Board of* Biopolymers *and on the Editorial Boards of* Proteins: Structure Function and Bioinformatics, Biosilico *and* IEEE Trends in Computational Biology and Bioinformatics. *He is the author of over 200 scientific papers and 4 books. He has received two UCSD Connect Awards for new inventions in the areas of comparative protein structure analysis and shared visualization. He was the recipient of the 2002 Sun Microsystems Convergence Award and the 2004 Convocation Medal for career achievement from his graduate university. He has co-founded four companies.*

Q: *What is the current impact of the PDB archive on biology, and what is the future of the archive?*

A: Given that more than six million data sets are downloaded from the wwPDB ftp archives each month, clearly its impact is large. The archive is recognized as a critical component in new drug discovery and development processes, and in the advancement of structural biology. While part of this usage is well understood–for example, there are many instances where structure provided a better understanding of biological function in disease states that led to the treatment of those diseases through new drugs–I suspect that there is a lot more to this story. A challenge in the next 5 years for all of the wwPDB is to better understand usage patterns and to help specific communities use the PDB archive in a way that would be the most beneficial to research and education.

Education is of keen interest at the RCSB PDB. Students in grades K-12 will be the leading scientists of tomorrow, and make up a key focus for our outreach programs. Structure biology has an advantage, as it is a visual science that can captivate young people. The RCSB PDB reaches out to these students through resources such as the New Jersey Science Olympiad and the *Molecule of the Month*. Of course, we would very much like to do more. One way we could proceed would be to take advantage of changing usage patterns on the web. Students today are very communicative online and use various social networking sites for hours on end. They are also part of the "Wiki Generation", where knowledge is defined by community input and consensus. Perhaps we at the RCSB PDB could capture this collective knowledge from teachers and students to create lessons around specific molecules and classes of molecules?

Q: *What about the older generation?*

A: For established life scientists, structure is often not a consideration and yet it has a great deal to offer. It is my experience that many life scientists associate molecular biology with DNA and protein sequences, and then skip structural biology to consider biochemical pathways, cellular processes, and whole cells and organisms. Let me give you an example from recent research work in my laboratory that makes this point using

evolutionary biologists as the test case. Since the time of Darwin, evolution has been studied through simple observation by paleontologists, zoologists, and botanists. Molecular biology, through protein and DNA sequencing, has revolutionized these evolutionary studies and allowed us to confirm and adjust the tree of life. But sequence has its limitations. The sequence signal degrades over long evolutionary time scales, and distant relationships cannot be seen. Structure is far more conserved than sequence over evolutionary time scales. With our ability to map structures to the ever-increasing number of fully completed proteomes, new insights can be made. Very few evolutionary biologists think of using structure in this way. One recent study from our laboratory showed how the tree of life could be reconstructed just by considering whether given species did or did not contain specific structural superfamilies of proteins defined by SCOP. In my view, the RCSB PDB has a role in facilitating these new kinds of studies to bring them to the attention of a broader community. So in this example, we could facilitate these studies by mapping structural domains and their changing arrangements onto the tree of life.

Q: ***Given these kinds of developments, where do you see the RCSB PDB in 10 years?***

A: The core mission of the RCSB PDB–providing timely delivery of high quality and complete structure data and useful and unique views of that data to enable scientific innovation–will not change. Of course, there will continue to be more and different types of data and the RCSB PDB will need to maintain these high standards of quality while catering to new types of delivery technology. It is hard to believe that the Internet has only been with us in a big way for ten years or so. Given the fundamental change in how we do science that has been bought about by the web, it is at least conceivable that how we do science will change even more dramatically in the future, even though we are hard-pressed to detail what those changes might be at this time. I would guess that we would need to provide data to people, software, and applications in seamless ways at very different degrees of granularity. Currently, most RCSB PDB queries return specific structures, but in the future you can imagine many more fine-grained requests from specific classes of scientist. For example, the pharmacy students I teach might use their handheld devices to ask a question like "we see significant instances of myocardial infarction in patients on select estrogen receptor modulator drugs like tamoxifen; what is the underlying biochemistry and molecular biology causing these side effects?" The RCSB PDB's role in this request could conceivably be to return and compare the receptors known to bind this class of drugs and allow the student to better understand the molecular implications. Inherent in this kind of request is the RCSB PDB's ability to integrate with other resources that permit the field of genomic medicine to advance and to return data such that non-specialists can answer their questions. These are significant (but fun) challenges.

Q: ***Let's bring you back more to the immediate future. The wwPDB recently remediated the entire PDB archive. What effect has this had on the RCSB PDB's query and reporting engine?***

A: The remediation effort is fundamental to the more far-reaching developments like those I have just discussed. Consistent representation of the data we have and the data we will collect going forward is critical if we are to use the archive effectively and integrate with other sources of data.

A very pragmatic example is the work that has gone into the Chemical Component Dictionary. As a result of this project, we can now reliably query ligands in the PDB archive through their names and/or chemical structures.

Q: ***You are heavily involved with the computational and systems biology community – how do these scientists use the RCSB PDB?***

I would say my area of work is best characterized as "structural bioinformatics," which is a small part of computational and systems biology. Even with this group of scientists where structure is central, the computational and systems biology communities are not yet taking full advantage of what we have. Most work is still performed using PDB files rather than XML files, and hence a lot of useful information is not being utilized by this community. This will change slowly as a generation of scientists more adept at dealing with XML start to have a stronger voice in the community. In terms of the bigger picture, systems biology is in some sense the molecular simulations of the new age. Rather than simulate the actions of a few molecules we are simulating the actions of complete pathways, cells and more. For now, at least, this work has largely bypassed structure, but I suspect that will change. The devil is in the details, and in the world of systems biology, structure may well provide those details. For example, much effort is currently going into mapping and describing the topologies of protein-protein interaction networks across a wide range of species and cell types. Eventually, it will be necessary to come back and explore specific interactions and here structure will be important. The challenge then for the RCSB PDB is to make available in a simple way the details of those interactions.

Q: ***You are very much involved with the* Public Library of Science (PLoS), *a nonprofit organization committed to making the world's scientific and medical literature freely available to the public. Why is that important to you?***

My work with PLoS is similar to our work on the PDB, where we all work hard to make data freely available to the worldwide community. PLoS tries to do the same thing, but with the scientific literature. PLoS is a standard bearer for the open access movement and I am very passionate about it. Open access to the literature is a controversial issue, and I appreciate the many sides to the argument (which I will not get into here). It is important to me that anyone can read the results of my research, but I acknowledge that open access is a business model that is yet to be proven. Nevertheless, I believe there is one component of open access that is very important. Open access is not just about access, but about copyright and format. Allowing anyone to use material from an article, provided they provide the appropriate attribution, opens up many possibilities when that information is marked up in XML and accessible online. My research group is experimenting with this through two NSF-funded pilot studies. The first is to integrate journal content with database content. Data and the knowledge derived from that data have traditionally been reported and kept separate (databases vs. publications). There is no reason for this, and so we are trying to come up with ways to provide more seamless and useful access between data and literature. This would seem to be particularly relevant to structure biology. A PDF file is a pretty poor way to express the aspects of a structure-function relationship that need to be looked at graphically. Take a simple example: a reader could go to a paper, and upon seeing a figure, click that figure. An identical version of that figure could be launched in a form that could be rotated, annotated, and used to ask for more information. My lab is developing prototypes that fulfill this idea. A second effort integrates open access content with video. We have developed a site called **scivee.tv** that attempts to cater for the upcoming YouTube generation of scientists. In "pubcasts," authors talk about their work in a video which is then integrated with the open access content of their paper. Relevant parts of the paper can highlighted as they speak. It remains to be seen whether scientists like this approach and whether it improves our ability to comprehend complex material. If the answer is "yes", it may be useful to include these kinds of developments into the RCSB PDB.

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. PHILIP E. BOURNE, Co-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 Center for Research Resources, the National Institute of Biomedical Imaging and Bioengineering, the National Institute of Neurological Disorders and Stroke, and the National Institute of Diabetes & Digestive & Kidney Diseases.*

RCSB PROTEIN DATA BANK
www.pdb.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