Winter 2006 • Number 28

RCSB PDB PROTEIN DATA BANK

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

NEWSLETTER

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

Contents

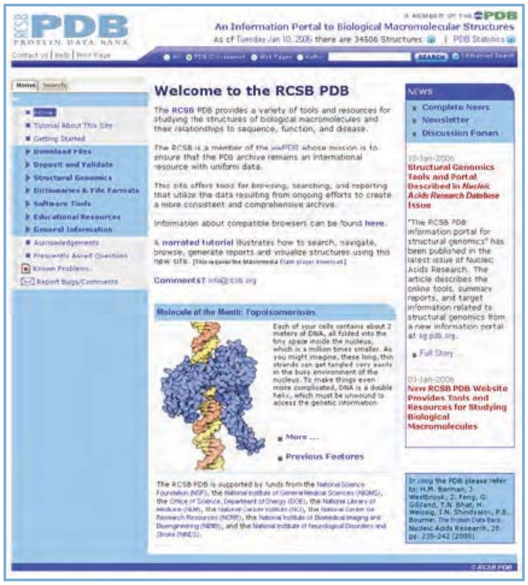
SNAPSHOT: JANUARY 1, 2006

34376 released atomic coordinate entries

MOLECULE TYPE		EXPERIMENTAL TECHNIQUE	
31414	proteins, peptides, and viruses	29211	diffraction and other
1543	nucleic acids	5165	NMR
1406	protein/nucleic acid complexes	19214	structure factor files
13	carbohydrates	2782	NMR restraint files

PARTICIPATING RCSB MEMBERS: RUTGERS • SDSC/UCSD
E-mail: info@rcsb.org
Web: www.pdb.org • FTP: ftp.rcsb.org
The RCSB PDB is a member of the wwPDB **(www.wwpdb.org)**

Message from the RCSB PDB



An Information Portal to Biological Macromolecular Structures

As of Tuesday Jan 10, 2006 there are 34506 Structures | PDB Statistics

Welcome to the RCSB PDB

The RCSB PDB provides a variety of tools and resources for studying the structures of biological macromolecules and their relationships to sequence, function, and disease.

The RCSB is a member of the wwPDB whose mission is to ensure that the PDB archive remains an international resource with uniform data.

This site offers tools for browsing, searching, and reporting that utilize the data resulting from ongoing efforts to create a more consistent and comprehensive archive.

Information about compatible browsers can be found **here**.

A **narrated tutorial** illustrates how to search, navigate, browse, generate reports and visualize structures using this new site. [This requires the Macromedia Flash player download.]

Comments? info@rcsb.org

Molecule of the Month: Topoisomerases

Each of your cells contains about 2 meters of DNA, all folded into the tiny space inside the nucleus, which is a million times smaller. As you might imagine, these long, thin strands can get tangled very easily in the busy environment of the nucleus. To make things even more complicated, DNA is a double helix, which must be unwound to access the genetic information.

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), and the National Institute of Neurological Disorders and Stroke (NINDS).

10-Jan-2006
Structural Genomics Tools and Portal Described in *Nucleic Acids Research Database* Issue

"The RCSB PDB information portal for structural genomics" has been published in the latest issue of Nucleic Acids Research. The article describes the online tools, summary reports, and target information related to structural genomics from a new information portal at sg.pdb.org.

03-Jan-2006
New RCSB PDB Website Provides Tools and Resources for Studying Biological Macromolecules

In citing the PDB please refer to: H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne. The Protein Data Bank. Nucleic Acids Research, 28 pp. 235-242 (2000).

At the very end of 2005, the RCSB PDB moved the updated version of the resource into production. The website and database at **www.pdb.org** have been enhanced and revised to provide a powerful portal for studying the structures of biological macromolecules and their relationships to sequence, function, and disease.

The RCSB PDB thanks everyone who has contributed to the development of this new resource since testing began in July 2004. Questions about the transition to this site not addressed in the FAQ should be sent to **info@rcsb.org**.

The new database utilizes PDB data that has been remediated and standardized for better searches and reports. Other features include improved ligand searching, a clear distinction between the reported primary and derived data, and the integration of external data resources (such as SCOP, CATH and chromosome location). A permanent search tab offers different ways of accessing the database, including a new method for "browsing" through structures grouped in categories (related to, for example, disease, molecular function, biochemical process, or cellular location).

Navigation of the website has been enhanced to keep the resources and tools related to structural genomics, education, and software within easy reach. A searchable help system with a glossary and user guide provides detailed information for accessing the website and database, and for understanding PDB data. A narrated presentation (in Flash) guides users through searching, navigating, generating reports, visualizing structures, and browsing PDB data on the new site. The structural genomics portal is enriched with target summary reports for centers worldwide, databases that track the progress of protein studies (TargetDB and PepcDB), and a tool to explore the distributions of functions found among structural genomics structures, PDB structures, genomes, and homology models. The legacy FTP structure will continue to be supported at: ftp://ftp.rcsb.org.

RCSB PDB:
www.pdb.org

Tutorial:
www.rcsb.org/pdbstatic/tutorials/tutorial.html

FAQ:
www.pdb.org/pdb/static.do?p=home/faq.html

Data Deposition and Processing

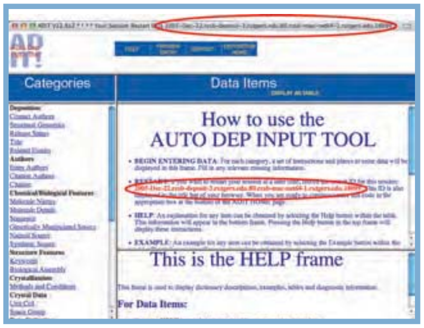
2005 Deposition Statistics

In 2005, 6491 experimentally-determined structures were deposited to the PDB archive – a 21% increase over 2004's 5356 depositions.

The entries were processed by wwPDB teams at RCSB-Rutgers, MSD-EBI, and PDBj. Of the structures deposited in 2005, 69.2% were deposited with a release status of "hold until publication"; 17.0% were released as soon as annotation of the entry was complete; and 13.8% were held until a particular date.

81.1% of these entries were determined by X-ray crystallographic methods, and 15.3% were determined by NMR methods. 80.7% of these depositions were deposited with experimental data.

PDB Focus: Restarting ADIT Depositions



The restart ID can be used to work on a deposition at a later time.

A structure can be deposited in more than one nternet session by using ADIT's "Session Restart ID" feature. This identifier appears in red in the center of the browser window when ADIT's "deposit" step is first started. It is also seen in the title of the browser throughout the deposition session.

The case-sensitive restart ID should be entered in the space provided on the ADIT home page to return to the undeposited entry. Any data entered in a category are stored every time the user selects the SAVE button. All entered data associated with a particular entry can be accessed using the restart ID until the "DEPOSIT NOW" button is selected, for up to six months after the session has been last updated.

ADIT is available at the RCSB PDB and PDBj. A tutorial guide to using ADIT is available in English and Japanese. Example "in progress" deposition sessions are available to practice learning how to use ADIT at **rcsb-deposit-demo-1.rutgers.edu**.

PDB Focus: First Time Depositors...

There are a few steps a depositor can take to make the process of depositing a structure to the PDB quick, easy, and accurate! This is an iterative process – if you encounter problems at a particular step, please make the correction(s) and go through the steps again. These resources are all linked from **deposit.rcsb.org**.

1. Use the **pdb_extract** Program Suite to extract information needed for deposition from output files produced by many structure determination applications.
2. Check your structure with the **Validation Suite and Server** to ensure that the data being deposited is accurate and reflects what you intend to submit.
3. Run **BLAST (at NCBI)** to compare your sequence to sequence database references. Any necessary corrections can then be made to your sequence and coordinates.
4. Use **Ligand Depot** to find the proper codes for existing ligands, to link to other entries with a particular ligand, and to search for substructures.
5. Deposit your structure using **ADIT**, using its editor to add any missing information to the deposition.

For a detailed packet of information about first-time deposition, including reprints about validation and Ligand Depot, please send your postal address to info@rcsb.org with the subject line "first time depositor packet".

PDB Focus: Ligand Depot – a Small Molecule Information Resource

Ligand Depot is a data warehouse that integrates databases, services, tools, and methods related to small molecules bound to macromolecules.

An important tool to use when depositing PDB structures, this resource can be used to find codes for existing ligands, to link to other entries with a particular ligand, and to search for substructures.

If a ligand related to a deposition is not in Ligand Depot, please email the chemical diagram, name, and formula to **deposit@rcsb.rutgers.edu.**

ligand-depot.rutgers.edu/
bioinformatics.oupjournals.org/cgi/content/abstract/20/13/2153

Ligand Depot: a data warehouse for ligands bound to macromolecules Zukang Feng, Li Chen, Himabindu Maddula, Ozgur Akcan, Rose Oughtred, Helen M. Berman, and John Westbrook. (2004) *Bioinformatics* **20**, pp. 2153-2155.

Data Query, Reporting, and Access

Searching and Browsing for PDB Structures

The Search Tab on the new RCSB PDB site offers many different ways of accessing the structures contained in the PDB archives.

Searching

The Advanced Search allows for simple queries based on structure summary items, keywords, structure/sequence features, ligands, biology & chemistry items, materials/methods, primary publication, and IDs (PDB, PubMed, Swiss-Prot, GenBank, PIR). These searches can be run in a variety of combinations to produce either very specific or very general results.

The Latest Release option displays images and summary information for structures added to the archives in the most recent update.

Forms are available to search specifically by sequence or ligand structure.

Unreleased structures can be searched by ID, title, authors, and sequence (when available).

The Queries tab shows all searches performed during the current web session. The results of these searches can also be retrieved.

Several simple text-based search options are also available on every page of the new site. Users can search the PDB archives using specific PDB IDs, keywords, or authors; run text searches on the static webpages; or search the archives and static webpages at the same time.

Browsing

Browsers are available to navigate structures using classifications from Gene Ontology, EC nomenclature, source organism, disease, genome, SCOP, and CATH. Users can explore each category's hierarchy, view the number of associated PDB structures, and search for specific related structures. For example, the Disease Browser can be used to look at diseases involving the nervous system, such as fragile X syndrome and Tay-Sachs Disease. Selecting a disease of interest (*e.g.*, Alzheimer Disease) will return all of the structures known to be associated with that disease.

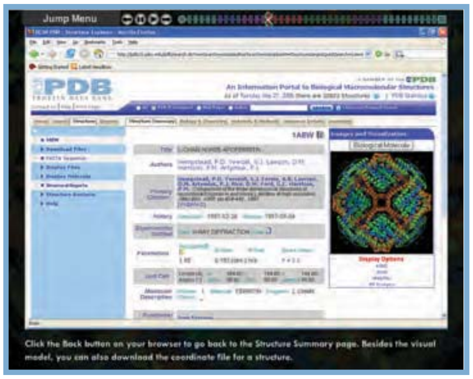
Reporting

For all searches, the resulting list of structures can be sorted, downloaded, used to create a tabular report (*e.g.*, citation or sequence information), or further refined by combining the search with another query.

PDB Focus: Help Systems for Searching the PDB, Depositing Structures, and More

Electronic help desks and an integrated help system are available to support users navigating the RCSB PDB.

The help system (accessible through each 🌐 button) launches into a separate browser window to access the help information and the website at the same time. It offers detailed topics (including Getting Started, Download Files, Search/Browse the Database, and Results), an index, a glossary, and a search engine.



Click the Back button on your browser to go back to the Structure Summary page. Besides the visual model, you can also download the coordinate file for a structure.

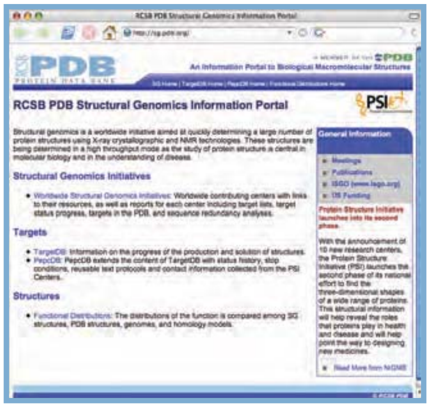
A narrated demonstration (shown on the left) is available to help users explore the new features of the website. This short tutorial provides an excellent overview for searching, navigating, generating reports, visualizing structures, and browsing PDB data.

deposit@rcsb.rutgers.edu answers questions about the deposition and annotation process at the RCSB PDB. Support pages at deposit.pdb.org include a file deposition and release FAQ, an overview of software tools, and tutorials for using ADIT, pdb_extract, the Validation Server, and Ligand Depot.

info@rcsb.org responds to requests relating to the navigation of the RCSB PDB website and database. Questions about searching, reporting, and using all of the resources available from the RCSB PDB should be sent to this address.

Enhanced Structural Genomics Portal

The RCSB PDB offers online tools, summary reports, and target information related to structural genomics at sg.pdb.org.


RCSB PDB Structural Genomics Information Portal

http://sg.pdb.org/

PDB
PROTEIN DATA BANK
A MEMBER OF THE PDB
An Information Portal to Biological Macromolecular Structures
SG Home | TargetDB Home | PepcDB Home | Functional Distributions Home
PSI

RCSB PDB Structural Genomics Information Portal

Structural genomics is a worldwide initiative aimed at quickly determining a large number of protein structures using X-ray crystallographic and NMR technologies. These structures are being determined in a high throughput mode as the study of protein structure is central in molecular biology and in the understanding of disease.

Structural Genomics Initiatives

- Worldwide Structural Genomics Initiatives: Worldwide contributing centers with links to their resources, as well as reports for each center including target lists, target status progress, targets in the PDB, and sequence redundancy analyses.

Targets

- TargetDB: Information on the progress of the production and solution of structures.
- PepcDB: PepcDB extends the content of TargetDB with status history, stop conditions, reusable text protocols and contact information collected from the PSI Centers.

Structures

- Functional Distributions: The distributions of the function is compared among SG structures, PDB structures, genomes, and homology models.

General Information

- Meetings
- Publications
- ISGO (www.isgo.org)
- DB Funding

Protein Structure Initiative launches into its second phase

With the announcement of 10 new research centers, the Protein Structure Initiative (PSI) launches the second phase of its national effort to find the three-dimensional shapes of a wide range of proteins. This structural information will help reveal the roles that proteins play in health and disease and will help point the way to designing new medicines.

- Read More from NIGMS

Information and links are provided for the worldwide structural genomics initiatives, including reports for each center that provide target lists, target status progress, targets in the PDB, and sequence redundancy analyses.

Databases that track the progress of protein studies are available. TargetDB contains information about the progress of the production and solution of structures. PepcDB extends the content of TargetDB with status history, stop conditions, reusable text protocols and contact information collected from the PSI Centers.

A tool is also provided to explore the distributions of functions found among structural genomics structures, PDB structures, genomes, and homology models. This functional coverage can be examined according to Enzyme Classification, Gene Ontology (Biological Process, Cell Component, or Molecular Function) and Disease.

A paper describing the methodology of this tool has been published: Lei Xie and Philip E. Bourne (2005) Functional Coverage of the Human Genome by Existing Structures, Structural Genomics Targets, and Homology Models. *PLoS Comput Biol* **1**(**3**): e31

compbiol.plosjournals.org/perlserv/?request=get-document&doi=10.1371/journal.pcbi.0010031.eor

Website Statistics

Access statistics are given below for the RCSB PDB website at www.pdb.org.

	DAILY AVERAGE		MONTHLY TOTALS			
MONTH	HITS	FILES	SITES	KBYTES	FILES	HITS
Dec 05	321,682	232,558	123,605	207,073,211	4,883,735	6,755,335
Nov 05	369,607	149,599	150,183	332,328,178	6,391,253	8,870,584
Oct 05	292,247	208,266	156,183	303,484,856	6,247,996	8,767,423

Outreach and Education

RCSB PDB's 2005 Annual Report Now Available

The RCSB Protein Data Bank's Annual Report, which covers the period of July 1, 2004 – June 30, 2005, is currently being distributed.

This snapshot of the RCSB PDB is intended to provide background information about the resource and describe recent progress and accomplishments. Available online as a PDF, this report describes the many different activities in data deposition, data access, and education and looks at the features of the new site.

If you would like a printed copy of the report, please send mail to **info@rcsb.org**.

www.rcsb.org/pdb/static.do?p=general_information/news_publications/index.html

Workshop for High School Teachers and Students: Building Protein Models at the Science Olympiad

"Protein modeling" – where students build physical 3D models of proteins – will be a trial event at the Northern Regional and State Final 2006 Science Olympiads in New Jersey.

This event will challenge high school students to explore structure and the relationship of structure to protein function using computer visualization and physical modeling tools. Students are introduced to the resources of the RCSB PDB, learn how to use RasMol, and create protein structures using 'Mini-Toobers' from 3D Molecular Designs.

The RCSB PDB team will be judging the models at the Olympiad and providing the materials for the event.

A training workshop was held at the RCSB PDB at Rutgers on Wednesday, December 7, 2005.

Tim Herman (MSOE Center for BioMolecular Modeling, **www.rpc.msoe.edu/cbm**) demonstrated how the Mini-Toobers are used to generate protein models in the competition.

Further details about this workshop and the olympiad event are available at **education.pdb.org/olympiad/**.

Teachers and students practiced making protein models from Mini-Toobers in preparation for the Science Olympiad.

RCSB PDB Exhibit, Workshop for New Jersey Science Teachers

Teachers examined the three-dimensional structure of biologically important macromolecules as part of the RCSB exhibit booth at the New Jersey Science Conference (October 5-6, 2005 in Somerset, NJ). The conference was co-sponsored by the New Jersey Science Teachers Association and the New Jersey Science Education Leadership Association. Two formal workshops were held in preparation for the protein modeling event at the Science Olympiad. At these meetings, Shuchismita Dutta presented "Seeing is believing but meeting is better" to introduce the educational resources of the RCSB PDB to science teachers. Gary Graper (Event Supervisor, Wisconsin Science Olympiad) and Jennifer Morris (Center for BioMolecular Modeling) gave hands-on demonstrations to show how the protein modeling event will work at the Science Olympiad.

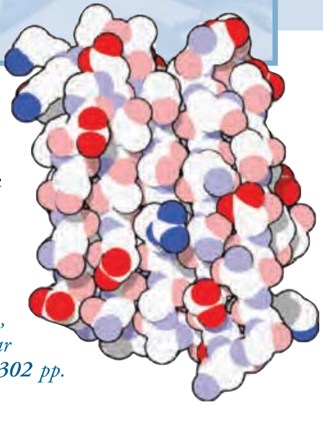
Molecules of the Quarter:

Designer Proteins, Acetylcholine Receptor, ATP Synthase

The **MOLECULE OF THE MONTH** *series explores the functions and significance of selected biological macromolecules for a general audience. The molecules featured this quarter were Designer Proteins, Acetylcholine Receptor, and ATP Synthase. The complete Molecule of the Month features are accessible from the RCSB PDB home page.*

Designer Proteins (October 2005)

As scientists began the quest to design entirely new proteins, they quickly found that proteins are more complicated than they might seem. The different types of amino acids, each with their own chemical features, work together to coax a protein chain to fold into a compact stable structure. A collection of carbon-rich amino acids, like leucine and phenylalanine, are usually placed inside the protein, all chosen to lock perfectly together. On the other hand, charged amino acids, such as lysine and aspartic acid, are typically spread across the surface to make the protein soluble in water. Hydrogen-bonding amino acids, such as serine and asparagine, are dotted in strategic places to tie different portions of the chain together. Finally, the odd glycine or proline is added to redirect the chain in the proper direction.



PDB ID: 1QYS

Kuhlman, B., Dantas, G., Ireton, G.C., Varani, G., Stoddard, B.L., Baker, D.: Design of a Novel Globular Protein Fold with Atomic-Level Accuracy Science ***v302*** *pp. 1364-1368, 2003*

PDB Education Corner:

Margaret A. Franzen, Pellissippi State Technical Community College

MARGARET FRANZEN *earned her Ph.D. in Biological Sciences from Northern Illinois University. Since 1997, she has been at Pellissippi State Technical Community College, where she teaches general biology, genetics and microbiology. In addition to covering course content, her teaching focuses on developing analytical thinking skills and exposing her students to the process of science. A number of her students have presented papers at collegiate meetings of the Tennessee Academy of Science. She recently won the Innovations in Teaching award at Pellissippi State and was a finalist in the Wiley* Life of the Classroom *innovative teaching competition in 2005. She has presented papers in state and national science education forums relating to the use of physical models and hands-on learning in the college classroom. Currently she is developing teaching materials for use with visually impaired, learning disabled and kinesthetic learners.*

Located in the beautiful Tennessee Valley between the Cumberland and Smoky Mountains in Knoxville, Tennessee, Pellissippi State serves over 7000 students on four campuses in both rural and urban settings. In addition to providing continuing education courses, the college offers two-year career and technical degrees as well as coursework for transfer to four-year institutions. Pellissippi State offers inexpensive educational opportunities to a diverse student body in small classrooms that permit excellent student-student and faculty-student interactions.

Background

I teach a sophomore-level college genetics course for biology majors as well as pre-vet and pre-med students. In the past few years, I have modified my syllabus to focus more on the relationship between the structure and function of proteins, as well as bioinformatics tools available on the internet. This article will briefly outline the activities I use in the classroom as well as the independent research projects that have developed as an offshoot of the course. The article will conclude with an overview of a bioinformatics activity that incorporates RasMol analysis of protein structure and physical models to study insecticide resistance that was developed as part of a National Science Foundation (NSF)-funded project.

A few years ago, the bulk of my genetics course was on Mendelian, transmission and population genetics; I emphasized regulation of gene expression, and only briefly addressed the nature of mutations. Students could define mutations but they never understood the relationship between the structure and function of a mutant protein.

In 2003 I participated in a summer workshop conducted at the Center for Biomolecular Modeling of the Milwaukee School of Engineering (CBM MSOE), where I discovered the value of a handheld physical model as a teaching tool. I also was exposed to RasMol, an elegant but compact computer program developed by Roger Sayle for manipulating protein coordinates.[1] Although RasMol doesn't have the 'point and click' features of some newer protein visualization programs, this program has the advantage of forcing students to *think* about what they are trying to do instead of simply pushing buttons to discover the effects on the visual appearance. Early exposure to RasMol also better equips our students for organic chemistry, which utilizes RasMol as a study tool.

Genetics Course

Early in the semester, I introduce my students to the Online Mendelian Inheritance in Man (OMIM) website.[2] As we discuss various genetic diseases or as they ask questions, I refer them to the website to discover the details of the diseases or the mutations that cause disease. By exploring, they learn that a mutant phenotype can often be caused by more than one mutation, sometimes involving completely different proteins. They also discover that mutations at different locations within the same gene can result in different diseases (*e.g.*, thalasemia and sickle cell anemia).

After students have learned the basic commands of RasMol, I have them pick a protein that is related to an altered human phenotype. Many of the students start at David S. Goodsell's Molecule of the Month features at the RCSB PDB site, where they learn about the normal functioning of the protein. They also utilize OMIM to learn more about genetic diseases related to their selected protein. Next, they are asked to design a model in RasMol that depicts the important features of the protein. This process forces the students to think about how the structure relates to the function of the protein. Students then post the PDB file, the RasMol script file they authored, and a paragraph describing the protein on a course discussion board within WebCT. The entire class is then able to visit the class 'protein gallery' to learn of multiple structure/function relationships.

Independent Study Projects

A number of students have taken a real interest in exploring protein structure in greater detail. These students have enrolled in an independent research course that is modeled after the SMART teams, described in an earlier Education Corner article (Spring 2003). The significant differences in the program are that the college students work individually rather than in teams to study a protein, and physical models are not always constructed at the completion of the project. Each student begins by locating the protein (often in various forms) in the PDB. After searching through the structure summary and sequence details of the various forms of the protein, students obtain copies of the original research papers as well as more recent articles by the primary investigators. Typically the depth of these articles is well beyond the background of the students, but they are able to make sense of the papers by using RasMol to highlight the features discussed. I encourage students to read the articles as though they were mastering a new language – by grasping the meaning of the unknown words from their context and immersing themselves in the papers. After a couple of readings of the papers in this fashion, students have grasped enough information to do a literature search on their protein, quickly ascertaining whether the articles are applicable to the features they want to study. I find that Google Scholar is a very useful tool for obtaining original articles. After digesting these articles, students are ready to interact, either directly or through email, with the original researchers. Although this step is not always possible (depending on the availability and interest of the researcher), it is incredibly beneficial for the students to discover 'how far they have come' in digging into a research topic. Next the students decide which features of the protein they want to demonstrate in their visual and, if possible, physical models. This requires a lot of playing around with RasMol and going back to the papers to verify details. At the conclusion of their work, students demonstrate their accumulated knowledge of the protein in a presentation at the collegiate meeting of the state science association. Students participating in the independent research project gain confidence in their ability to tackle 'real' science papers and communicate with others. Most are interested in science as a career before enrolling in the course, but are turned on to the possibility of doing research as a result of the program.

Insecticide Resistance Activity

This exercise was developed in conjunction with Dr. Tim Herman (CBM MSOE) and Dr. David S. Goodsell (The Scripps Research Institute) as part of an NSF Undergraduate Education (DUE) grant (#0442409). The activity incorporates both bioinformatics and analysis of protein structure (using PDB files) to determine the nature of insecticide resistance in mosquitoes. Many insecticides target the enzyme acetylcholinesterase, which breaks down the neurotransmitter acetylcholine in cholinergic junctions. Students repeat the work of Weill et al.[3] in determining the differences in acetylcholinesterase isolated from insecticide sensitive (S) and resistant (R) strains of the mosquito *Culex pipiens.* First they compare the DNA sequences of S and R strains, and then translate the sequences and align the protein sequences. Students discover a number of silent mutations, typically due to changes in the third position of the codons. There is only one amino acid difference (a Gly to Ser mutation) between the two strains! Optionally, students can also align a series of twenty-nine S and R strains (all posted at NCBI) to show that there is only one mutation that is consistently found in all resistant strains and none of the sensitive strains. Students then access acetylcholinesterase structures from the PDB to determine the location of the amino acid change. Working with RasMol, Jmol images, and physical models, students determine how the mutant enzyme can break down the substrate yet does not respond to inhibitor. One student's response to the exercise was, "I never realized that one carbon could make such a big difference." This activity is currently being developed as a Waksman Challenge for high school students, and physical models will be available from the MSOE Model Lending Library in the near future. This exercise allows students to relate much of what they have learned in the course of the semester to a practical application, thus reinforcing their understanding while also exposing them to useful research tools available on the internet.

A physical model of the acetylcholinesterase active site from CBM, with removable substrate and inhibitor molecules, as well as interchangeable glycine and serine residues for insecticide sensitive and resistant enzymes, respectively.

Links

Center for BioMolecular Modeling: www.rpc.msoe.edu/cbm

Google scholar: scholar.google.com

MSOE Model Lending Library: www.rpc.msoe.edu/lib

OMIM: www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM

Pellissippi State Technical Community College: www.pstcc.edu

RasMol: www.rasmol.org

Waksman Challenge: wakschallenge.rutgers.edu

WebCT: WebCT.com

References

1. Sayle, R. and Milner-White, E.J. RasMol: biomolecular graphics for all. *Trends Biochem. Sci.,* 1995. **20**: p. 374.

2. Wheeler, D.L., et al., Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res,* 2005. **33** (Database issue): p. D39-45.

3. Weill, M., et al., Comparative genomics: Insecticide resistance in mosquito vectors. *Nature,* 2003. **423**: p. 136-137.

PDB Community Focus:

Frank Allen, Cambridge Crystallographic Data Centre

Q: *The CCDC turned forty this year. You have been involved for 35 of those 40 years – how have the CCDC and the CSD changed during that time?*

A: The CSD has changed from being an academic idea to being an information resource that is used, and relied upon, by several thousand scientists in academia and industry. Annual CSD input has increased by a factor of 30, but amazingly electronic input has only taken over during the last decade. More than 200,000 of our current entries were re-typed from journals and deposition documents right up to the mid-1990s, but the CIF has now changed all that (for the better – although not all CIFs are perfect!). Chemically and crystallographically, we now process much larger structures, more of which are truly novel, and

FRANK ALLEN *was born in Reading, UK in 1944 and studied chemistry at Imperial College London, receiving a BSc in 1965 and a PhD in 1968. Following postdoctoral work at the University of British Columbia, Vancouver, he joined the University of Cambridge in 1970 to carry out small-molecule structure determinations. Work on the Cambridge Structural Database (CSD) had begun in 1965, and his interest in this project grew during the early 1970s. He has been involved in most major developments at the Cambridge Crystallographic Data Centre* ***(www.ccdc.cam.ac.uk)*** *since then, including software development and database creation, but with a strong accent on research applications of the CSD. He received the Royal Society of Chemistry Prize for Structural Chemistry in 1994 and the Herman Skolnik Award of the American Chemical Society Division of Chemical Information in 2003. He is now Executive Director of the CCDC and a Visiting Professor of Chemistry at the University of Bristol.*

many more of which are metal-organic species with challenging problems of structure representation.

The most fundamental organizational change occurred in 1989, when the CCDC became financially self-supporting and self-managing. This has been both a challenge and a benefit, and one that has enabled us to develop carefully as the business has grown. From a maximum of about 15-20 staff in its agency-funded days, the CCDC is now a nonprofit company employing 50 people. Those staff are now highly focused, with software being developed, released and supported to professional standards, and increasing automation being brought to bear on the creation, validation and maintenance of the CSD itself.

Q: ***The CSD now contains about 370,000 published structures. However, far more structures than that have actually been determined but are not published, maybe approaching a million. Is there any way in which some or all of these structures can be collected into the CSD?***

A: This is indeed a serious problem. Very large numbers of high-quality small-molecule structures are lying dormant in the filing cabinets and internal archives of hundreds of laboratories across the world. The reasons for this situation are varied, and are discussed in an article I wrote recently in *Crystallography Reviews* (vol. 10, 3-15, 2004). Principally, service crystallographers perform small-molecule structures for chemists, give them the results, and that is as far as many of them get. The original structural problem is solved and synthetic or other work can go ahead (or not!). It is really a question of 'ownership' of the crystal structure data, and the will (and time) to place them in the literature, an open archive or a database. Every crystal structure is valuable, and often for reasons that are unrelated to the original research goals. So, it is a major problem for structure-based science that so much valuable data is well on the way to being lost forever. This may also become an issue for macromolecular structures as well, and it is to be hoped that some of the developments noted below may mitigate the problem. Chemoinformatics and bioinformatics approaches to a whole host of structural problems depend on fully comprehensive (and accurate) data resources. These informatics approaches cannot be maximally effective if the flow of available data is attenuated, as it clearly is now for small molecules.

The CCDC has made strong efforts to attract unpublished data into the CSD, and more than 2,000 unpublished structures have been deposited directly into the CSD in the past 5 years. Section E of *Acta Cryst.* has also benefited the community by publishing several thousand structures which might otherwise not have seen the light of day. However these numbers are a drop in the ocean, and the CCDC has supported e-Science initiatives at Southampton, UK and Indiana, USA to encourage unpublished structural data into the public domain. We also welcome proposals announced by the NIH and the UK Research Councils for public archiving of research results obtained with their funding. We must wait to see exactly how these e-Science and funding agency initiatives develop, and will be talking to the organizers of the archives to see how we can be involved, so as to maximize the value of the CSD to the scientific community.

Q: ***How have your collaborations affected the CCDC?***

A: Wholly positively. The CCDC has collaborated with many institutions on both the research applications of the CSD, and in the co-development of products. Research collaborations enable CCDC staff to be involved in novel uses of the CSD and suggest improvements to the distributed software that arise from the work. A recent example is the CCDC's involvement as a partner with Cambridge University and Pfizer Inc. in the Pfizer Institute for Pharmaceutical Materials Science. This has brought together experimental, computational and chemoinformatics approaches to address typical problems faced by development and formulation chemists in the pharmaceutical industry. The research is suggesting new ways of organizing and searching CSD data, and is giving rise to novel software to accomplish these aims. On the product development side, we have co-developed the protein-ligand docking program GOLD with GlaxoSmithKline and the University of Sheffield, Relibase and Relibase+ with Merck, Germany and the University of Marburg, and the DASH software for structure solution from powder diffraction data with the Rutherford Appleton Laboratory in the UK. This has not only broadened the scientific interests of the CCDC, it has also generated additional income which has helped both the CCDC and its academic collaborators, and has enabled us to keep subscriber costs for the CSD itself as low as possible over the past decade.

Q: ***It would be of great value to users of both the RCSB PDB and the CSD if there was a direct two-way connection between the two databases. What do you foresee as possible collaborations between the CSD and the RCSB PDB and when do you think it could happen?***

A: There are a number of ways in which the RCSB PDB and the CSD could work together. It seems to me that the best way forward is for the CCDC and the RCSB PDB to discuss possibilities together and then bring forward some joint proposals, taking account of their scientific value and the available resources at both organizations. Only then can we determine a way forward and realistic timescales. I don't think it wise for either organization to make unilateral statements.

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, UCSD
9500 Gilman Drive
La Jolla, CA 92093-0537

The RCSB PDB is a member of the WORLDWIDE wwPDB PROTEIN DATA BANK (www.wwpdb.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, and the National Institute of Neurological Disorders and Stroke.

RCSB Leadership Team

The overall operation of the PDB is managed by the RCSB PDB Leadership Team. Technical and scientific support is provided by the RCSB PDB Members.

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

DR. PHILIP E. BOURNE, Co-Director
San Diego Supercomputer Center
University of California, San Diego
bourne@sdsc.edu

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

JUDITH L. FLIPPEN-ANDERSON, Outreach Coordinator
Rutgers, The State University of New Jersey
flippen@rcsb.rutgers.edu

DR. JOHN WESTBROOK, Co-Director
Rutgers, The State University of New Jersey
jwest@rcsb.rutgers.edu

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

RCSB PROTEIN DATA BANK
www.pdb.org
9650 Rockville Pike
Bethesda, MD 20814
Return Service Requested