Bio·IT World
INFORMATION TECHNOLOGY FOR THE LIFE SCIENCES
October 2002 • VOL. 1, NO. 8

Banking on Structures

An explosion of structural information is on the horizon, and the Protein Data Bank — the single international repository for data on the three-dimensional structures of biomolecules — is ready.

BY TRACY SMITH SCHMIDT

Except for the small sign at the end of the walkway, you wouldn't guess that the nondescript, low-level brick building on the campus of Rutgers University is home to one of the world's most important biological databases — the Protein Data Bank (PDB). It seems too insignificant to contain all the personnel and equipment needed to manage the current flow of information into and out of the database, let alone handle the approaching surge of structural data.

Indeed, it is too small. The Rutgers location in Piscataway, N.J., is only one of three sites responsible for the PDB, along with the San Diego Supercomputer Center (SDSC) at the University of California, and the National Institute of Standards and Technology (NIST) in Gaithersburg, Md. Collectively, they make up the Research Collaboratory for Structural Bioinformatics (RCSB). Rutgers is responsible for data processing, SDSC for managing the public database, and NIST for maintaining the physical archive of PDB files. Together, they appear well-equipped to handle as much structural data as the community can throw at them.

The PDB harbors the coordinate files that pinpoint the location of nearly every atom in thousands of proteins and nucleic acids — maps obtained using structural techniques such as X-ray crystallography and NMR (nuclear magnetic resonance) spectroscopy (see "Solving Structures," page 3). Imaging programs can interpret these files, allowing scientists to visualize the convoluted shape of a protein and assess how it may interact with carcinogens or drugs.

One of the oldest biological databases, the PDB was founded in 1971 in a very forward-thinking move, considering that only about a dozen protein structures had been solved at the time and just a few scientists required access to the data. But by the time the RCSB replaced Brookhaven National Laboratory as custodian of the database in 1998, the PDB contained approximately 8,000 entries. That number has now grown to more than 18,000. Today, the demand for structural data permeates every biological field — from antibiotic resistance in bacteria to learning and memory in humans. The database is free to all on the Web (www.pdb.org) and has become an indispensable resource for biologists of all stripes.

PDB for Free

The Protein Data Bank is free to search on the Web at www.pdb.org. Biologists can deposit data, download files, beta-test new features, and check the status of unreleased protein structures at the site.

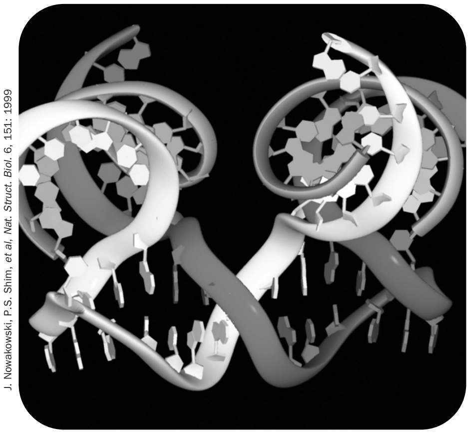
The amount of data in the PDB should grow rapidly in the near future thanks to high-throughput structural genomics efforts underway in both academia and industry. The goal of those efforts is to produce thousands of structures per year, adding to the PDB's current annual intake of about 2,000 to 3,000 structures. (See "Betting on the Structural Revolution," September *Bio•IT World*, p. 60.)

"It's coming," says Helen Berman, director of the PDB, who has prepared for just such an onslaught by automating procedures and modernizing the database structure. Berman still maintains a laboratory at Rutgers, but feels that guiding the PDB into the new era of structural genomics is a contribution to science that her own research program cannot match. "I believe this is the most important thing for me to be doing," she says. "I would like people to know the level of seriousness with which we treat the data."

Even though the results housed in the PDB reflect years of work, surprisingly, the entire data set takes up a mere 20GB to 25GB — a quantity that would easily fit on a single laptop. "It is not so much the sheer amount of data as it is the complexity," says Wolfgang Bluhm, production manager at the SDSC site. Bluhm is one of some 30 staffers distributed across the three sites, although some telecommute using a Linux platform from places as far-flung as Prague in the Czech Republic.

Data Processing

The PDB has the formidable task of formatting, annotating, validating, and releasing dozens of complicated structure files every week. "Data processing is the heart of the PDB," says Kyle Burkhardt, senior biochemical information specialist on the team. Depending on the size of the protein, it can take anywhere from a couple of hours to several days. Without painstaking attention to every file, the archive would lose uniformity and, with it, query power. The PDB has automated as much of this operation as possible to prepare for the surge in submissions.



J. Nowakowski, P.S. Shim, *et al. Nat. Struct. Biol.* 6, 151; 1999

Crystal structure of an 82-nucleotide RNA-DNA complex formed by the 10-23 DNA enzyme, which is stored on the PDB.

What is the path of a typical PDB file? First, a researcher uploads the data to the private Rutgers deposition server and receives a PDBid, a unique four-character code for the structure. Next, a PDB employee manually combs through the information, looking for missing pieces of data, and also runs validation software to catch anomalies, such as unusual bond angles in the structural model. The annotated file is then returned to the scientist for revision. Once the depositor and the PDB staff have approved the file, it is ready for release via a weekly update sent to the SDSC site, where the public PDB database is managed (see figure below).

"Determining the proper balance between what human beings should do and what computers can do has been difficult," says John Westbrook, co-director of the PDB and head of the Rutgers site. Nevertheless, the team has found a good equilibrium, because total processing time averages only two weeks, and the annotators keep the backlog as low as possible — only 50 to 100 files at worst.

"The most time-consuming part is validating the structure," says Westbrook. Three Compaq AlphaServers (models ES40 and ES45) are dedicated to this task. The PDB has made its validation suite of software available on its Web site, and in May it began distributing the source code in the hope of expediting the entire process.

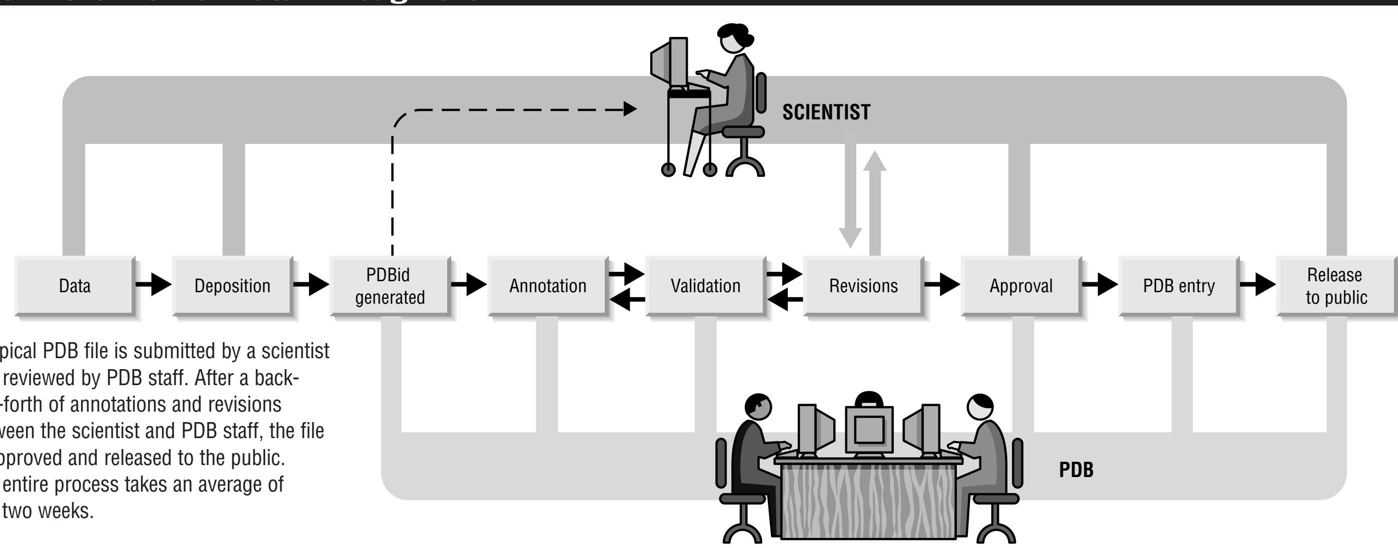
Harvest Festival

A concept known as data harvesting — whereby all of the information necessary for the PDB file is collected automatically, minimizing human error — could reduce processing time even further. There are nine federally funded structural genomics consortia in the United States, and each is currently developing harvesting procedures.

Tom Terwilliger, a scientist at Los Alamos National Laboratory and the Structural Genomics Consortium's principal investigator of Mycobacterium tuberculosis, defines data harvesting as "trying to help the PDB collect as much information about structures as possible, as quickly as possible." Others, such as Bob Sweet, a scientist at Brookhaven, see it primarily as a labor-saving device and insurance against omitting data or introducing errors.

There is currently no set of best practices for collecting and standardizing all the data. Both the PDB and the National Institute of General Medical Sciences (NIGMS), which funds the structural genomics consortia, have been encouraging these efforts through workshops. "Their hope is that the best ideas will sort of percolate and develop," says Ray Stevens, a scientist at The Scripps Research Institute, who also founded structural genomics company Syrrx Inc. and is a core leader in the Joint Center for Structural Genomics. Some groups have begun testing their harvesting protocols with the help of the PDB.

Steps in the Flow of Data Through the PDB



A typical PDB file is submitted by a scientist and reviewed by PDB staff. After a back-and-forth of annotations and revisions between the scientist and PDB staff, the file is approved and released to the public. The entire process takes an average of just two weeks.

ANNIE BISSETT

PDB Politics

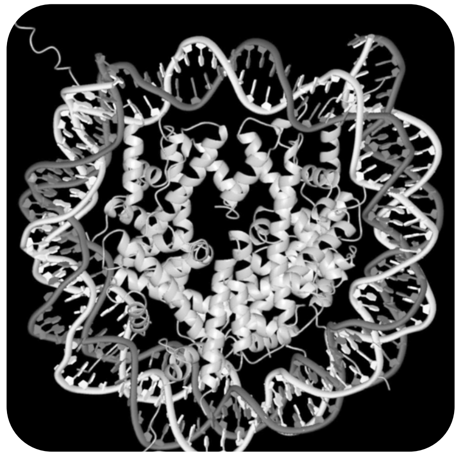
From day one, the PDB has been freely accessible to scientists around the world, but funding comes solely from the United States — pieced together from the National Science Foundation, the Department of Energy, the NIGMS, and the National Library of Medicine.

Still, the PDB does have strong ties to overseas scientific organizations. The European Bioinformatics Institute (EBI) in the United Kingdom and the Institute for Protein Research in Japan both serve as extra PDB deposition and data processing sites. The procedures used at the Institute for Protein Research mirror those at Rutgers — in fact, PDB personnel trained their staff. The EBI uses different software but provides the final files in a format suitable for exchange with the PDB's preferred style. Both groups send their finished files to the PDB for inclusion in the definitive archive.

There are distinct advantages to having multiple deposition sites. It spreads out the workload (Rutgers alone receives about 40 files per week), and scientists can easily communicate with the local processing staff. However, some scientists worry that, without agreement, such a setup could lead to the Balkanization of the PDB. Berman insists that this is not likely to occur, since cooperation is now close and tight. "We did it right — we held the line on a single archive," she says, referring to the widespread consensus on preserving the PDB as the single clear repository for structural data. Though each location may provide different and useful ways to interpret the information, all of the raw data will be stored in one definitive database: the PDB.

Banking everything in a unified format, however, is easier said than done. When the RCSB took charge, the PDB was a set of inconsistently annotated flat files in various states of repair, with no formal data representation as found in modern relational databases. To renovate the system, the RCSB adopted mmCIF (macromolecular crystallographic information file) as the format for describing the structural data in the PDB.

According to Westbrook, "The mmCIF dictionaries were 10 years in the making — figuring out what should be included and writing them out" in consultation with scientists to guarantee that all of the important information from a crystallographic experiment would be recorded. The product is a searchable dictionary table comprising thousands of terms. The RCSB has since migrated some 8,000 "legacy" files from Brookhaven into mmCIF — in some cases retrieving data from decades-old tapes to make the archive as complete as possible.



K. Luger, A.W. Mader, et al, Nature 389, 251: 1997

Crystal structure of the nucleosome core particle at 2.8 A resolution, from the PDB.

The mmCIF format is highly extensible. "Everything we do is based on an electronic model that is scalable, that can grow," says Westbrook. To encapsulate additional data, the RCSB simply adds another set of terms. For example, structural genomics researchers are eager to store comprehensive information about the expression, purification, and handling methods leading to the final structures. "The PDB has done a great job," says Terwilliger. "The idea is to have all of the data that led to the structure in the PDB so that people can look into it further, do additional experiments."

In principle, the PDB can house all sorts of information, including proteomics and protein-folding results, as long as suitable dictionaries are available. However, Berman stresses that the team's current mandate is to preserve only information relevant to experimentally derived structures. Consistent with this, they recently removed all theoretical structural models from the main archive, making them available instead as a collection on an FTP site. Although they will continue to accept computationally derived models, they will not curate them with the same care afforded experimental data.

Solving Structures

PDB keeping up with protein data that come in all shapes and sizes

Proteins are the workhorses of the cell, performing nearly every function required for life. A protein chain typically folds into a specific structure, ready to perform a certain task, such as digesting sugar or carrying oxygen in the blood. So to understand how a protein works, it is essential to discover how it folds into a particular conformation.

Deciphering molecular structures requires indirect approaches to tease out information about the positions of the atoms in a molecule. Two of the most commonly used techniques are X-ray crystallography and NMR (nuclear magnetic resonance) spectroscopy. In both cases, the researcher, with the help of numerous computer programs, pieces together a model of the molecule that best fits the experimental data.

In X-ray crystallography, researchers shine X-rays on crystals that contain trillions of copies of a molecule. The crystals diffract the rays in patterns that, when analyzed mathematically, reveal the positions of the atoms within the molecules.

NMR spectroscopy uses molecules in solution rather than in a crystal, and relies on the innate property of atoms to orient themselves in a magnetic field. In this situation, atoms produce distinct "signatures" that give clues to their environments within the molecule. About 15 percent of the structures in the Protein Data Bank (PDB) were solved using NMR.

These procedures result in a plethora of information — how the protein was extracted and purified, how it was prepared (crystallized or concentrated in solution), as well as the raw data from the experiments themselves that reveal the atomic coordinates. While both techniques yield maps of atom locations within the molecule, there are many differences in the data leading to the final coordinate file.

The mmCIF (macromolecular crystallographic information file) format — the set of data dictionaries used in the PDB's relational database — was developed specifically for crystallographic results, but additional dictionaries are in the works for NMR. "NMR data are quite a bit more complex than X-ray data, much richer and more difficult to capture," says the University of Wisconsin's John Markley, who also heads the NMR database for the BioMagResBank (BMRB).

The PDB and the BMRB teams are working together to ensure that the important data are archived efficiently; they are also developing a common deposition tool so that the appropriate data go to the right database. **— T.S.S.**

Database Developments

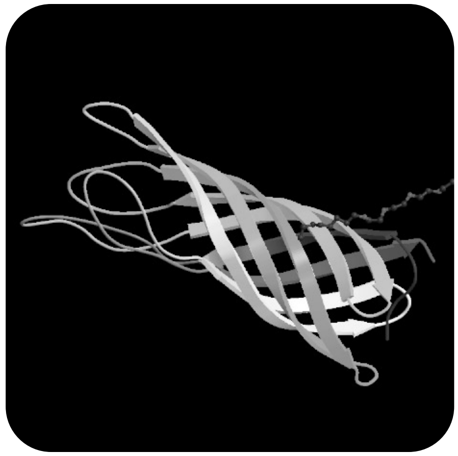
One of the reasons the PDB archive offers such a complete record of experimental structural biology is that funding agencies and scientific journals require the deposition of relevant data as a condition of financial support or publication. Structural results cannot be fully appreciated in two dimensions, so access to the primary coordinate files is crucial. There are also tremendous benefits to searching across structures and linking to analysis tools available on the Internet — features provided by the PDB.

Currently, the PDB uses a Sybase database to store the primary data. This component is integrated via Perl Common Gateway Interface (CGI) scripts with several other resources, including the Netscape Lightweight Directory Access Protocol (LDAP) server, which supports keyword searches, and the Biological Macromolecule Crystallization Database, which contains literature-derived information on crystallization procedures.

Plans are under way to unveil a new version of the database soon, using IBM's DB2 platform. "The goal is alpha testing this fall and beta testing in the spring," says Berman.

"IBM has been extremely helpful," says Phil Bourne, co-director of the PDB and head of the SDSC site. "They are even providing a DB2 expert on site [at the SDSC] for several months to get us started."

This second-generation database will be more accessible to independent programmers and bioinformaticians, as well as to other databases that regularly run scripts to hunt for relevant links. The Object Management Group, which oversees the development of specifications for life science computing, recently accepted the Common Object Request Broker Architecture (CORBA) application programming interface (API) that had been proposed by the PDB to define biomolecular structure data. The new database will allow use of this CORBA-based API and a Java-based API, giving users a choice when writing their software.



A. Pautsch, G.E. Schulz, *Nat. Struct. Biol.* 5, 1013: 1998

The structure of the outer membrane protein A (OMPA) transmembrane domain.

"We would like to make a significant contribution to making a 'database federation,'" says Bourne, referring to the recognized need in the community to better integrate scientific databases. Adding new APIs is a clear step in that direction.

One of the biggest challenges faced by the PDB is to maximize the potential of its Web site to serve a diverse set of constituencies — hard-core structural biologists, molecular biologists who dabble in structural biology, graduate students, and even high school students. Currently the site receives over 100,000 hits per day, with one structure downloaded on average each second, 24 hours a day.

The new Web interface will provide more visuals, site maps, and streamlined navigation, as well as other enhancements. For example, Bourne says the now uniform data set and database structure will permit more complex queries. "A user will be able to look at the structure of, say, an ATP-binding protein, click on the ATP ligand [in the image], and find out what other structures bind ATP."

But Bourne adds that the development of querying capability is limited by the facts that depositors provide. "There is a point of balance between providing information and endorsing specific ways of interpreting the data," he says. For example, the PDB does not sanction any single way of doing sequence alignments because not all scientists agree on the most appropriate methodology. The PDB consequently offers links to various methods and resources.

A program developed at the SDSC called the Molecular Information Agent automatically searches for PDBid codes in about 75 external resources and creates reciprocal links to those pages in the appropriate PDB files. For example, in PDB file 1CDW, which is the structure of a human DNA-binding protein, a user can find links to the Structural Classification of Proteins database and the Columbia Picture Gallery, to name only two.

On the Horizon

The RCSB team has undoubtedly worked hard to improve the PDB. "I am proud that we have been able to meet all the goals we said we would, with no one having a nervous breakdown," says Berman, smiling.

But they are not sitting back. "Our short-term goals are to get the second-generation database into alpha testing, to finish the materials and methods data dictionaries, and to encourage people to use the software we have made available," says Berman. In the longer term, the team will deal with the impact of structural genomics, look for ways to improve database interoperability, and try to contribute to the development of international funding solutions for databases.

As genomics moves to the next level — analyzing the entire complement of proteins in numerous organisms, including humans — the value of the PDB data will only increase. Berman's team is ready for the challenge. "We are acutely aware of thinking about what our scope is, and the scope changes all the time," says Berman. "All that we do is community-driven. We listen to what people want and ask ourselves how we can change our system to give it to them." ●

Tracy Smith Schmidt is a writer based in New York City and the former editor of Nature Structural Biology. *She can be reached at tracy_schmidt@nasw.org.*

The Nucleic Acid Database "Incubator"

Early adapter to relational database approach helped set the standard for revamped Protein Data Bank

The Protein Data Bank (PDB) is not the first database that Director Helen Berman has run. She and several other members of the Research Collaboratory for Structural Bioinformatics (RCSB) team cut their teeth on the Nucleic Acid Database (NDB), which was established in 1991.

Although DNA, RNA, and protein-nucleic acid structures have always been welcome in the PDB (despite the noninclusive nature of its name), the NDB team realized early on that a relational database would be of greater benefit to the nucleic acid research community than the flat files found in the PDB. So they set out to build one. The NDB vision was a well-curated database of primary structural results and derivative data that would allow complex queries and comparisons of nucleic acid structures.

The NDB team tested the mmCIF (macromolecular crystallographic information file) format and found that it worked well. They introduced structure validation software and automated data processing procedures, and added resource links to their Web site to help scientists. In essence, the NDB experience made the RCSB highly qualified to take over the larger, more complex PDB in 1998. The NDB's software and systems were easily transferable to the PDB — most of the major bugs had already been eliminated.

Berman still oversees the NDB, which, unlike the PDB, is funded by a research grant that enables it to act as an incubator of ideas. Berman is cautious about altering the PDB archive but can play around a little more with the NDB.

To come up with new ways of doing things, biological databases need those rare individuals with experience in two languages — science and computers. This past summer, the NDB rooms were full to overflowing with such a group — undergraduate interns with double majors in biology and computer science, who were eagerly collaborating on new tools for the site. Berman clearly approves: "I wish I knew how to bottle this recipe so that we'd have it every year." **— T.S.S.**