The Protein Data Bank and the challenge of structural genomics

Helen M. Berman1, T. N. Bhat2, Philip E. Bourne3,4, Zukang Feng1, Gary Gilliland2,5, Helge Weissig2 and John Westbrook1

The PDB has created systems for the processing, exchange, query, and distribution of data that will enable many aspects of high throughput structural genomics.

The determination of structures on a genomic scale in a high-throughput model will have an impact on every aspect of the Protein Data Bank (PDB) — the single archive for all biomacromolecular structural data. Although estimates vary, the PDB could triple in size over the next five years2–4. Not only is it likely that the number of structures will increase dramatically, but the information about each structure is destined to grow as well. Uniform experimental practices will offer the opportunity to automatically collect more data from each structure determination. The quality of structures from high-throughput experiments may be more variable, depending upon such factors as the extent of refinement. Now, in addition to archiving the results of structural biology projects driven by the need to answer questions arising from the results of biochemical experiments, we will need to catalogue structures for which little or no functional information is available2. How will the PDB respond to changes in quantity, quality and available functional information in this new era?

In 1998, the Research Collaboratory for Structural Bioinformatics (RCSB; http://www.rcsb.org), a consortium consisting of Rutgers University, the San Diego Supercomputer Center (SDSC) at the University of California, San Diego (UCSD), and the National Institute of Standards and Technology (NIST) became the custodians of the PDB2. The PDB was established in 1971 with a handful of structures at Brookhaven National Laboratory2. The RCSB has provided new integrated systems for collecting, distributing and querying the contents of the PDB. These systems were designed with the expectation that there would be a change in the quantity, quality, and content of new structures. It is the goal of the PDB to enable new scientific research through effective use of structure data; however, the provision of the data may change. The PDB will actively cooperate with emerging structural genomics centers in the public and private sector to make the best quality data available in a timely manner.

To understand how the PDB is responding to structural genomics, we consider the steps involved in processing and distributing a structure, and how these will expand and be modified in this new era of structural biology.

Data deposition and processing

The PDB captures and processes data using the AutoDep Input Tool (ADIT; Fig. 1). The data submitted via this Web interface include atomic coordinates, journal references, information about the biological macromolecule (such as the name, sequence, and source), ligand information, experimental information, and primary data (such as structure factors and NMR constraints). During data processing, formats and nomenclatures are standardized and many different checks are made on the structure, agreement of the model with the primary data, and sequence. The time spent during the checking process is a function of the complexity of the molecule.

All of the software used for the collection and processing of input data is driven from the macromolecular crystallographic information files (mmCIF) dictionary. The mmCIF dictionary, created under the aegis of the International Union of Crystallography (IUCr)5, contains 1,700 terms that rigorously define the various aspects of the crystallographic experiment and the resultant biological structure. This computer-readable and parsable dictionary also defines data types, enumerates ranges of allowable values where possible and describes allowable relationships between data values5. This structure enables the creation of relational databases and exact translation into other formats such as XML. In order to continue to allow the dictionary to expand as the science changes, formal procedures, again under the auspices of the IUCr, were developed to create new data items. New widely used data items are reviewed by a board of editors and incorporated into the dictionary. Recent examples of additions relate to newly established features of structure refinement.

The dictionary is extensible and can be used to define terms for other types of experiments such as NMR and cryo-electron microscopy. This characteristic of the dictionary design allows for customization and will be very important as new data items will have to be developed for the structural genomics effort. mmCIF

---

1Department of Chemistry, Rutgers, the State University of New Jersey, 610 Taylor Road, Piscataway, New Jersey 08854-8087, USA. 2Biotechnology Division, Chemical and Science Technology Laboratory, National Institute of Standards and Technology, Gaithersburg, Maryland 20850, USA. 3San Diego Supercomputer Center, University of California, San Diego, 9500 Gilman Drive, La Jolla, California 92037-0537, USA. 4Department of Pharmacology, University of California, San Diego, 9500 Gilman Drive, La Jolla, California 92037-0500, USA. 5Center for Advanced Research in Biotechnology (CARB), National Institute of Standards and Technology and the University of Maryland Biotechnology Institute, 9600 Gudelsky Drive, Rockville, Maryland 20850, USA. Correspondence should be addressed to H.M.B.

email: berman@rcsb.rutgers.edu
Data exchange

The worldwide structural genomics efforts require the free exchange of data among the participants and the PDB. Conformance to a common standard is clearly desirable if that exchange is to be complete and exact. An effect of conformance to the mmCIF dictionary is to provide the appropriate level of data exchange so data are easily interpreted and data are not lost in the exchange. The PDB format, while the de facto standard, is clearly not rigorous or complete enough to provide this guaranteed level of data exchange. The PDB and the EBI-MSD have chosen mmCIF as the exchange format and have recently completed finalizing the terms that will be used by the EBI when they send processed data to the PDB each week. The PDB will begin releasing expanded data in mmCIF format by the end of 2000. The PDB will also assist by the provision of basic software tools, in the conversion of mmCIF to other formats and standards.

Data quality

New data collected as part of the structural genomics projects will necessarily be compared and integrated with the data that is part of the current archive. It is thus essential that the existing

Fig. 1 ADIT, the integrated deposition and processing system, is available at http://pdb.rutgers.edu/adit/. Current system is in gray; future enhancements are in gold. The client input tool collects data from the output of programs involved in various aspects of structure determination. These include programs involved in data collection (beamline control), data processing programs, refinement programs, beamline profiles (BioSync), and crystallization conditions. ADIT is built upon metadata dictionaries, such as mmCIF, and the data processing program called the Macromolecular Exchange Input Tool (MAXIT)\(^1\). Other tools employed by ADIT include the validation checks run on the data, the database loader, and the flexible data views.

Fig. 2 The integrated PDB Query System, is available at http://www.rcsb.org/pdbi. Current system is in grey; future enhancements are in gold. A user query (from SearchLite or SearchFields) is transparently passed on to one or more of several underlying databases optimized to efficiently return different types of data (Query Result Browser and Structure Explorer pages). In response to structural genomics, an interface for use by other databases and applications will be added, as well as new query capabilities and analysis tools.

also provides external reference files (ERFs) which contain standard data such as nomenclature and standard geometries.

The presence of a framework for defining terms, the existence of ERFs, and the flexibility of specialized views allow ADIT to be readily used for the high-throughput demands of structural genomics. To be most efficient, the information required for PDB deposition should be collected automatically during the course of the experiment. This concept, called data harvesting or data pipelining, is already being implemented. Working in collaboration with the European Bioinformatics Institute (EBI), the Collaborative Computational Project (CCP4)\(^{10}\) for protein crystallography will produce mmCIF data files that may be automatically uploaded to the PDB. PDB staff members are also participating in an analogous project, the Collaborative Computational Project for NMR (CCPN), which is just being initiated with the NMR community (http://www.bio.cam.ac.uk/nmr/ccp/). In a joint project between members of the PDB and BioSync (http://biosync.sdsc.edu), a US synchrotron user’s group, all of the attributes of the beamlines are resident as mmCIFs. HKL2000 (ref. 11) creates mmCIF compliant files as does CNS\(^2\) and X-PLOR\(^3\). As a proof of concept, in November 1999, data were collected at the “High Throughput Methods for Structural Genomics” meeting held at Argonne National Laboratory and rapidly deposited into the PDB using this pipelining approach.

To facilitate fast and accurate data deposition activities critical to high-throughput structural genomics, the PDB will undertake the following:

(i) Assist in the creation of data definitions for the new types of data items that need to be archived automatically. These include details of target selection, protein expression, purification, characterization, crystallization, data collection, structure determination, and refinement that are not yet part of the mmCIF dictionary. This work will be done in collaboration with an informatics task force that was created at the First International Meeting on Structural Genomics held in April 2000 in Hinxton, UK.

(ii) Work with software developers in the integration of these new data items into their software and make sure the resultant data files can be automatically uploaded to the PDB.

(iii) Make available PDB data processing and validation software so that structure depositions can be checked prior to deposition, thus saving time and improving the quality of the depositions. In addition, the PDB will incorporate into its checking procedures the numerical criteria used for evaluating structure quality that are being developed by a task force created for that purpose at Hinxton.
information content be reliable and uniform with respect to nomenclature and format. The PDB has undertaken a major effort to ensure this uniformity. This is done in two ways: file-by-file and batch. In the batch method individual pieces of information, for example R-factor, resolution, and primary citation are checked across the entire archive, often by returning to the original literature, and standardized. In addition, synonyms for protein and ligand names are added to tables in the database. A file-by-file analysis is even more rigorous across all structures in a given protein family. While some uniform data are already available when accessing the PDB, the results of this work need to be made generally available. After consultation with the community it was clear that the original PDB files should not be modified and re-released. Rather, the results of all of this work will be put into a single ftp site with structures presented as mmCIF formatted files. For those preferring PDB formatted files they can be created from this site using tools developed by the PDB.

**Data distribution and query**

The increased quantity of data expected from structural genomics will have little impact on the data distribution of PDB data since the process of loading the databases and distribution to mirror sites worldwide is automated. This process was designed to scale to a much larger number of entries than are currently being processed on a weekly basis. Data quality may be evaluated using the currently available validation server ([http://pdb.rutgers.edu/validate/](http://pdb.rutgers.edu/validate/)) as well as new stereochemistry and experimental data checking tools that will be made available on the PDB web sites.

Structural genomics will lead to a variety of new web accessible resources containing information relevant to structures found in the PDB. In the PDB’s role as an Internet portal these resources should be accessible immediately as a user obtains a structure of interest. The PDB achieves this level of linkage automatically using the same technology developed for the Molecular Information Agent (MIA; [http://mia.sdsc.edu/](http://mia.sdsc.edu/)). The PDB-specific implementation of MIA is a web crawler that accesses several predefined resources, queries them for PDB-related information and automatically creates a link to the remote information specific for a PDB entry. Automatic, periodic updates insure a current set of web links to related sources of information worldwide. Currently MIA accesses 60 resources, but this number is expected to increase to encompass sites such as NIH-funded structural genomics center web sites.

While the PDB provides many links to related information at remote sites, it is important in this new era that the PDB provide more seamless and direct access to its data for remote users and applications. The current WWW interfaces employ the standard Common Gateway Interface (CGI) protocol (Fig. 2). CGI allows remote users to execute programs at a web resource that provides access to query functionality and tools implemented by the information resource itself. In contrast, technologies like the Common Object Request Broker Architecture (CORBA) provides direct access to information for programs designed and written by remote users.

CORBA provides a standard Interface Definition Language (IDL) for describing distributed data resources like the PDB. Software tools can be used to convert this CORBA IDL into a language-specific application programming interface (API). This API can be used by remote users to implement new query and analysis tools and to exchange data directly with any resource that has implemented a server conforming to the same API.

The PDB has submitted a proposal describing the specifications for a macromolecular structure IDL based on the mmCIF data representation. The proposal is currently under consideration by the Object Management Group (OMG), the organization that oversees and standardizes the CORBA IDLs. IDL specifications are also being reviewed for genomic maps and biomedical sequence analysis. Collectively, these specifications should provide a robust framework for the interoperability of key data resources required by the structural genomics effort. In simple terms, a specification will exist for a remote application to seamlessly retrieve a single item of data from the PDB for use as part of a local application. At present, it is only possible for a user to retrieve the complete PDB entry (see [http://www.rcsb.org/pdb/linking.html](http://www.rcsb.org/pdb/linking.html)).

Efforts are now under way to annotate the raw sequence comprising the human genome. In the future we anticipate the large-scale annotation of PDB data for which no functional description was originally assigned at the time of structure determination. The integration of the PDB with other resources of information will be critical in providing scientists with a complete view of a macromolecular structure. How such annotation will be validated and how, and if, it is made part of the PDB entry remains a challenge to the PDB and will be decided through community consensus. An immediate goal is to provide annotation that indicates structures that have been derived as part of the structural genomics initiative.

**Acknowledgments**

The PDB is supported by funds from the National Science Foundation, the Office of Biology and Environmental Research at the Department of Energy, and two units of the National Institutes of Health: the National Institute of General Medical Sciences and the National Library of Medicine.

**Associations with structural genomics**

The authors are members of the Research Collaboratory for Structural Bioinformatics. G.G. is associated with Structure 2 Function ([http://s2f.carb.nist.gov/](http://s2f.carb.nist.gov/)) a collaboration between CARB, the TIGR Institute for Genomic Research, and the Illinois Institute of Technology. P.B. is a participant in the Joint Center for Structural Genomics at The Scripps Research Institute and a consultant for Structural GenomiX.