Acta Cryst. (2002). A**58** (Supplement), C214

DATA ANNOTATION, PROCESSING METHODS, AND TOOLS AT THE PDB AND NDB

Z. Feng[1,2] H. Hyang[1,2] J. D. Westbrook[1,2]
[1]Protein Data Bank Rutgers, The State University of New Jersey Department of Chemistry 610 Taylor Road PISCATAWAY NJ 08854-8087 USA [2]Nucleic Acid Database

The Protein Data Bank (PDB; **http://www.pdb.org/**) is the international repository for three-dimensional structural data. The Nucleic Acid Database (NDB; **http://ndbserver.rutgers.edu**) specializes in nucleic acid-containing structures. These resources have developed several integrated tools for the deposition and query of structures in these databases.

ADIT (AutoDep Input Tool) is used by the community for data validation and deposition, and internally by the PDB and NDB for processing and annotation. Because it is based on the data dictionary technology of mmCIF, ADIT is easily extended to accept and process information describing new science and technology. Examples of how this system is used for processing some of the most complex structures in the PDB and NDB will be described.

The ADIT system has been designed to operate as a distributed network application or as a stand-alone tool. A workstation version of ADIT is available at **http://deposit.pdb.org/software/** for researchers to prepare and validate entries in their home laboratories. This system also includes a variety of utility programs to assist in the extraction of information from several crystallographic applications, and utilities for merging the various program outputs into a single mmCIF data file ready for PDB deposition.

The PDB is managed by three members of the RCSB: Rutgers, NIST, and SDSC. The PDB project is funded by the NSF, DOE, and two units of the NIH: NIGMS and NLM. The NDB is funded by the NSF and the DOE.

Keywords: DATABASES, STRUCTURAL BIOINFORMATICS, MMCIF

Acta Cryst. (2002). A**58** (Supplement), C214

THE PROTEIN DATA BANK: DATA DISTRIBUTION AND QUERY FUNCTIONALITY

W.F. Bluhm[1] T. Battistuz[1] E. Clingman[1] N. Deshpande[1] W. Fleri[1] D. S. Greer[1] D. Padilla[1] D. Stoner[1] H. Weissig[1] P.E. Bourne[1,2,3]
[1]Protein Data Bank San Diego Supercomputer Center University of California, San Diego 9500 Gilman Drive SAN DIEGO CA 92093-0537 USA
[2]Department of Pharmacology, University of California, San Diego [3]The Burnham Institute, La Jolla, CA

The Protein Data Bank (PDB; **http://www.pdb.org/**) is the international repository for the processing and distribution of three-dimensional structural data. Its mission is to enable science by providing the most accurate and timely data for macromolecular structures. Data distribution and query functionality are replicated at six additional mirror sites, each of which maintains a Web site and an ftp archive. Weekly data updates are first tested on a local staging site, and then distributed to all production sites. All new functionality is first released on a public β test site (**http://beta.rcsb.org/pdb/**) prior to its distribution to all production sites. Examples of added query or display functionality include an enzyme classification browser, customized tabular reports, the pre-release of sequence information for some unreleased structures, and the STING Millennium Suite of graphical structure/sequence viewing tools (courtesy Goran Neshich and Barry Honig). Since the PDB holdings contain a considerable amount of redundancy, a sequence homology filter was implemented that provides the choice of displaying either a representative set of structures or the full search results. Progress on a re-engineering effort of the database, software, and Web interface will also be described. The PDB is managed by three members of the RCSB: Rutgers, NIST, and SDSC. The PDB project is funded by the NSF, DOE, and two units of the NIH: NIGMS and NLM.

Keywords: DATABASES, DATA DISTRIBUTION, QUERY FUNCTIONALITY

Acta Cryst. (2002). A**58** (Supplement), C214

CREATING THE CAMBRIDGE STRUCTURAL DATABASE(CSD): A CCDC EDITORS VIEW OF DATA ACQUISITION, VALIDATION AND REGISTRATION OF NEW ENTRIES

K. L Foreman
CCDC 12 Union Road CAMBRIDGE CB2 1EZ UK

The CSD now records over 260,000 small-molecule crystal structures and more than 24,000 new structures were added during 2001. About 85% of new datasets now arrive at the CCDC in CIF format via the Internet prior to publication, and the CCDC is the official data depository for more than 50 major journals. Individual datasets from this depository are provided free of charge. Once data are matched with a publication, a CSD entry is constructed from both deposited material and publication by adding or improving chemical information, such as a formal 2-D diagram, compound name(s), formula, etc. The entry is then validated and further refined by, inter alia, checking that published geometry and coordinates are self-consistent, that the chemical and crystallographic connectivities agree, that the chemical and crystal data are self-consistent, and that disorder is handled according to current CSD conventions. Problems are solved in-house or in collaboration with original authors, often leading to text comment being added to the entry. Entries are then registered against the master CSD, to detect additional determinations of existing structures; CSD reference codes are adjusted and the entry is archived. Work is now under way to generate more Editor-friendly software systems for data acquisition, processing and registration, and new systems will enhance the data content of CSD entries using the deposited CIF data as the basis.

Keywords: CAMBRIDGE STRUCTURAL DATABASE, DATABASE CREATION, DATA VALIDATION

Acta Cryst. (2002). A**58** (Supplement), C214

NEW CIF EDITING SOFTWARE AND EXAMPLES OF COMMON CIF PROBLEMS

B. Smith C.F. Macrae O. Johnson
Cambridge Crystallographic Data Centre 12 Union Road CAMBRIDGE CB2 1EZ UK

The Crystallographic Information File (CIF) is now a standard means of information exchange in crystallography. An increasing number of journals generate papers for publication directly from CIFs, and it is also the recommended way of submitting data to the Cambridge Structural Database (CSD).

At the Cambridge Crystallographic Data Centre (CCDC) we are developing new CIF editing software for use by the scientific community as well as our own database staff. One of our main aims is to create a program that provides an intuitive, user-friendly interface to enable you, having completed a crystal structure refinement, to add information safely to the resultant CIF file without corrupting the strict syntax. As well as catering for users who are familiar with the CIF format, much attention has been placed on designing features suitable for those with less experience. Features include: a publication wizard to help you generate a file with all the necessary data, a powerful loop editor, visualization of three-dimensional structures, on-line help for data items, syntax checking and validation.

The presentation will also focus on some common CIF problems, such as those we regularly encounter whilst processing data submitted to the CSD. The meeting should also be an opportunity for us to gain feedback from potential users. Our software is currently at the stage of β testing. It will be available as a free download for Windows and Unix/Linux platforms.

Keywords: CIF VALIDATION SOFTWARE