DOE/ER/62761

MACROMOLECULAR STRUCTURE DATABASE

DE-AI02-~~01~~99ER62761

FINAL PROGRESS REPORT

Gary L. Gilliland

NIST

Structural Biology Group

Biotechnology Division

Chemical and Science Technology Laboratory

National Institute of Standards and Technology

DISCLAIMER

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

The central activity of the PDB continues to be the collection, archiving and distribution of high quality structural data to the scientific community on a timely basis. The systems we have helped to develop for doing this have become increasingly reliable and stable. In support of these activities NIST has continued its roles in developing the physical archive, in developing data uniformity, in dealing with NMR issues and in the distribution of PDB data through CD-ROMs.

Physical Archive Development. The physical archive contains the paper documents, magnetic tapes, and other materials generated by the PDB over the course of its history. The database and automated filing system, developed for the paper archive of the structures as well as the data read in from old tapes, continue to be called on periodically to resolve issues involving older structure records. The reading of older backup tapes is continuing. To accommodate the restored data, 1TB of disk storage has been installed on the computer systems being used to house the electronic archive. This produces many copies of the same file. As data are restored, they are incorporated into a file structure that is designed to trace their origin, that aids in retrieving information, and that can be easily maintained and expanded.

A Physical Archive website has been developed for the PDB staff to facilitate access to the electronic archive and inventories. Efforts are underway to establish a unique (non-redundant) file set. In order to enable the staff to find data more quickly within the archives, software that compares the files and reports unique sets is being tested. As the data files are moved onto the external archive system they are being written to new media as much of the older media will become unreadable. Tapes now unreadable locally are sent to a specialist in reading tapes. In addition, copies of each release of CD-ROM/documentation/flyers, Newsletters, Annual Reports are added to the PDB Physical Archive. The disaster recovery tapes for the PDB website are duplicated and added to the archive; the copies of these tapes are stored in a secure location in a separate building.

With the reorganization of the physical archive materials, the next phase of working with the physical archive involves scanning and electronically storing documents associated with the PDB. A diverse set of paper files associated with depositions were scanned and are under review to establish which of the documents should be made electronically available for the PDB staff's use. The files include both correspondence about individual entries and administrative records.

The recovery of electronic files from the legacy magnetic media will be carried out until all have been read. Software is being developed for creating a non-redundant fileset from the hundreds of backup tapes comparing will be incorporated into the electronic archive and used to facilitate the searching of the electronic archive by PDB staff members. Methods for complete indexing of the files by PDB ID and author have been under development to enhance the usefulness of this resource.

Data Uniformity. The efforts on data uniformity for entry data items on macromolecular name, source, citation, etc. have continued. Recently, the data for x-ray and NMR data collection and processing and the software and hardware used in structure analysis have been evaluated and annotated. This process involved establishing standard names and synonyms for data elements. These data elements have been incorporated into scalable data definitions to facilitate interoperability and use by other members of the RCSB and user community. The work to date

has been incorporated into the Protein Data Bank entries and made available to the public for evaluation. This has been done by incorporating the results of these and other efforts of the PDB into a set of mmCIF files that also contain all of the original author-submitted information. These data will be released initially in a beta-data ftp site.

NMR. We have continued to collaborate with the NMR community to establish a standard data representation for PDB entries. The PDB has also undertaken a collaborative effort with the BioMagResBank (BMRB) to develop a joint deposition system based on PDB's ADIT software. The goal of this collaboration is to provide a single integrated deposition system for NMR data that will accept both experimental and structure data. A preliminary version of an ADIT server for the data items collected by the BMRB is expected to result from this effort next year. This version will use an mmCIF-like dictionary (NMRIF) that was derived from the NMR STAR deposition form used by the BMRB.

CD-ROM Production. CD-ROM production has occurred quarterly without fail. The October 2002 release of the PDB CD-ROM set (Issue 102) included the 18,796 structures released as of October 1, 2002. Five CD-ROMs were required to contain these structures in compressed (gzip) format. With this release, the experimental data files (NMR constraints and X-ray structure factors) are available as separate products. The x-ray structure factors require five CD-ROMs and the NMR constraints require one. Subscribers were notified of these changes with the July release, by PDB news articles, and with the October release so that they were given an additional opportunity to order the experimental data if they required it. With the October release of CD-ROM set, the theoretical models were put into a *models* directory, separate from the experimental coordinates. CD-ROM production will continue on a quarterly schedule. Software was developed to allow incremental distribution of the CD-ROM data. Starting with the January 2003 release, a full release of data (coordinates and experimental) will be issued in the first quarter, followed by quarterly updates. New subscribers after the first quarter will receive the January release as well as all of the incremental sets up to that date.