Databases III-25 Years of PDB

MS22.03.01 OVERVIEW & PLANS FOR THE BROOKHAVEN PROTEIN DATA BANK. J.L. Sussman, Brookhaven National Laboratory, P.O. Box 5000, Upton, NY, 11973-5000 USA and Dept. of Structural Biology, Weizmann Institute of Science, 76100 Rehovot, Israel.

The Protein Data Bank (PDB) at Brookhaven National Laboratory (BNL) is an archive of experimentally determined three dimensional structures of biological macromolecules, serving a global community of researchers, educators, and students. The challenges presented by the enormous growth in data over the past several years have been met by the PDB staff and management, which now provides an up-to-date archive, while simultaneously expanding network access (Stampf, Felder & Sussman, *Nature* **374**, 572-574, 1995) and building strong world-wide collaborations (Peitsch, Stampf, Wells & Sussman, *TIBS* **20**, 82-84, 1995). Even larger challenges are at hand, as the deposition rate continues to rise along with the expectations of the consumers of these data. We are enhancing the capabilities of the PDB and transforming it into a new *Three-Dimensional Biomacromolecular Structure Database*, while retaining many of the features of the PDB, for compatibility.

Over the course of the next few years, we plan to:

- Ensure that the archive remains current, correct, and relevant.
- Provide a rapid and painless automated deposition system, using internationally agreed-upon CIF standards.
- Enhance data validation tools.
- Build a relational database to store and provide flexible access to the information.
- Integrate the PDB with complementary biological and chemical databases through semantic links and schema sharing.
- Provide easy access, through the Internet, to the PDB at BNL and other deposition and distribution sites worldwide.

The PDB is supported by a combination of Federal Government Agency funds and user fees. Support is provided by the U.S. National Science Foundation, the U.S. Public Health Service, National Institutes of Health, National Center for Research Resources, National Institutes of General Medical Sciences, National Library of Medicine, and the U.S. Department of Energy under contract DE-AC02-76CH00016.

MS22.03.02 CREATION OF THE PROTEIN DATA BANK. Edgar Meyer, Biographics Laboratory, Dept. of Biochemistry and Biophysics, Texas A&M University, College Station, Tx 77843-2128 USA.

A library has three principal characteristics: Location, holdings, and users. The years 1968-1971 encompass the stages of conception, gestation, and delivery of the PDB. While the PDB is fundamentally scientific in content, the thesis will be supported that technological factors were primary determinants which influenced these principal characteristics. The events and decisions leading up to the creation and utilization of the PDB will be presented from the vantage point of past, present, and future technologies.

MS22.03.03 EARLY HISTORY OF THE PROTEIN DATA BANK AND THE EVOLUTION OF THE NUCLEIC ACID DATABASE. Helen M. Berman, Department of Chemistry, Rutgers University, Piscataway NJ 08855.

In 1971, a handful of protein crystal structures had been studied to a point where atomic coordinates had been determined. Yet it was already apparent to a small group of crystallographers that a repository was needed for these valuable data so that they would be available to all interested researchers. Thus was born the Protein Data Bank (PDB). This talk will describe the early years of the PDB, the vision of the founding leadership, the involvement of the community, and the technology that was used to develop the archive.

The Nucleic Acid Database (NDB) was created in 1991 in order to provide a way to curate a special subset of the macromolecules contained within the PDB. In 1996 it became a direct deposition site for nucleic acids. The features of the NDB will be described as well as the procedures that have been put in place to coordinate with the PDB in the maintenance of the central macromolecular archive.

Support for this work has been provided by the NSF (BIR 9305135).

MS22.03.04 25 YEARS OF THE PDB: IMPACT ON STRUCTURAL BIOLOGY. David R. Davies, Gerson H. Cohen, Laboratory of Molecular Biology, NIDDK, NIH, Bethesda MD 20892

The current explosive expansion of structural biology has profited greatly from the existence of the PDB. In methodology, ready access to coordinates has facilitated the development and application of phase determining procedures such as molecular replacement. Analysis of the PDB has also provided statistical criteria for structural evaluation.

However, it is the area of comparative protein structure analysis that has been most affected by the flood of new structures and the availability of coordinates. Although many protein superfamilies (eg immunoglobulins, serine proteases etc) are related by similar functional properties and by sequence homologies, others (TIM barrels, 4 alpha helix bundles) are not and structural analysis has revealed hitherto unsuspected relations between proteins that are unsupported by functional or sequence similarities. Sophisticated procedures have been developed to analyse the relatedness of members of these families. As the number of protein entries in the PDB continues to rise, estimates can be made of the total number of independent protein folds which appears to be finite. A continuing goal will be to provide a basis for protein structure prediction based on amino acid sequence.

Blundell T. L. and Johnson M. S. Protein Science, 1993, 2:877
Murzin A. G., Brenner S. E., Hubbard T. and Chothia C. J Mol Biol, 1995, 247:536
Orengo C. A., Flores T. P., Taylor W. R. and Thornton J. M. Protein Engineering, 1993, 6:485.