

The Protein Data Bank: Current Status and Future Challenges

Joel L. Sussman^{1,2}, Enrique E. Abola¹, Nancy O. Manning¹ and Jaime Prilusky^{1,3}

¹Protein Data Bank, Biology Department, Brookhaven National Laboratory, Upton NY 11973-5000, USA

²Department of Structural Biology, Weizmann Institute of Science, Rehovot 76100, Israel

³Bioinformatics Unit, Weizmann Institute of Science, Rehovot 76100, Israel

jls@bnl.gov; <http://www.weizmann.ac.il/~jsgrp/joel.html>

abola1@bnl.gov

oeder@bnl.gov

lsprilus@weizmann.weizmann.ac.il; http://bioinformatics.weizmann.ac.il/jaime_prilusky.html

Abstract

The Protein Data Bank (PDB) is an archive of experimentally determined, three-dimensional structures of proteins, nucleic acids, and other biological macromolecules. The PDB is now being transformed into 3DB, the Three-Dimensional Database of Biomacromolecular Structures, with significantly enhanced capabilities.

Development is underway for 3DB to operate as a direct-deposition archive, providing mechanisms for depositors to submit data over the Internet with minimal staff intervention. Data archived in 3DB is managed using the Relational Database Management System (RDBMS), SYBASE [4]. The new database (3DBase) is being developed with a view to being a member of a federation of biological databases. Collaborative international centers are also being established to assist in data deposition, archiving, and distribution activities.

1 Introduction

The Protein Data Bank (PDB) is an archive of experimentally determined, three-dimensional structures of proteins, nucleic acids, and other biological macromolecules [1, 2]. PDB has a 25-year history of service to a global community of researchers, educators, and students in a variety of scientific disciplines [3]. The common interest shared by this community is a need to access information that can relate the biological functions of macromolecules to their three-dimensional structures. The PDB is now being transformed into 3DB, the Three-Dimensional Database of Biomacromolecular Structures, which will continue to operate from Brookhaven National Laboratory.

The challenge facing the new 3DB is to keep abreast of the increasing flow of data, to maintain the archive as error-free as possible, and to organize and present the stored information in ways that facilitate data retrieval, knowledge exploration, and hypothesis testing, without interrupting current services. The PDB has introduced substantial enhancements to both data management and archive access in the past two years, and is well on the way to converting to a more powerful system that combines the advantages of object-oriented and relational database systems. 3DB will transform PDB from a Data Bank serving solely as a data repository into a highly sophisticated knowledge-based system for archiving and accessing structural information. The process will be evolutionary, insulating users from drastic changes, and will provide both a high degree of compatibility with existing software and a consistent user interface for casual browsers.

2 Resource Status, September 1996

Rapid developments in preparation of crystals of macromolecules and in experimental techniques for structure analysis and refinement have led to a revolution in Structural Biology. These factors have contributed significantly to an enormous increase in the number of laboratories performing structural studies of macromolecules to atomic resolution and the number of such studies per lab. Advances include: 1) recombinant DNA techniques that permit almost any protein or nucleic acid to be produced in large amounts; 2) better X-ray detectors; 3) real-time interactive computer graphics systems, together with more automated methods for structure determination and refinement; 4) synchrotron radiation, allowing the use of extremely tiny crystals, Multiple Wavelength Anomalous Dispersion (MAD) phasing, and time-resolved studies via Laue techniques; 5) NMR methods permitting structure determination of macromolecules in solution; and 6) electron microscopy (EM) techniques, for obtaining high-resolution structures.

These dramatic advances produced an abrupt transition from the linear growth of 15-25 new structures deposited per year in the PDB before 1987 to a rapid exponential growth reaching the current rate of ~25 deposits per week (Fig. 1).

This rapid increase overwhelmed PDB staff resources and data processing procedures and, by mid-1993, a backlog of some 800 coordinate entries had accumulated. This backlog was eliminated by January 1994 by increasing automation of processing and hiring additional staff. In all, more than 4,250 of the ~5,350

current PDB coordinate entries (~80%) have been processed since 1991.

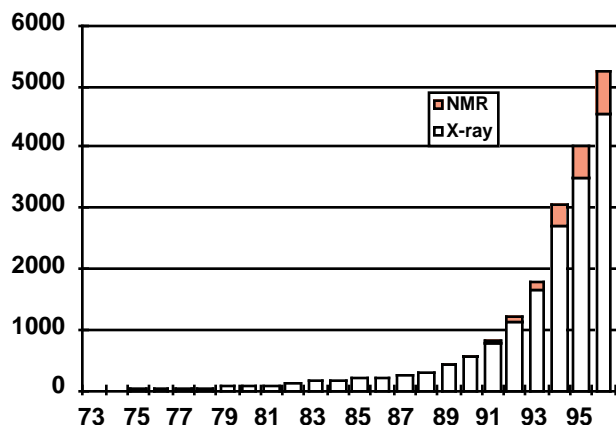


Figure 1 PDB Coordinate Entries Available

Table 1 is a summary of the contents of PDB. Present plans are to keep abreast of the deposition rate with a timeline of three months from receipt to final archiving, which includes the time that the entry is with the depositor for checking. This timeline is comparable to the publication schedules of the fastest scientific journals.

PDB Archive Contents	
5346	released atomic coordinate entries
576	structure factor files
243	NMR restraint files
Molecule type	
4748	proteins, peptides, and viruses
213	protein/nucleic acid complexes
373	nucleic acids
12	carbohydrates
Experimental Technique	
4441	diffraction and other
758	NMR
147	theoretical modeling

Table 1 PDB Archive Contents - Jan 1997

In the same period, the proliferation and increasing power of computers, the introduction of relatively inexpensive interactive graphics, and the growth of computer networks greatly increased the demand for access to PDB data (Fig. 2). The requirements of molecular biologists, drug designers, and others in academia and industry are often fundamentally different from those of crystallographers and computational chemists, who have been the major users of the PDB since the 1970s.

PDB entries are available on CD-ROM, which PC users can search using the PDB-SHELL [5] browser, built using FoxPro [6]. In addition to its browsing

mechanisms, PDB-SHELL provided direct access to the molecular viewing program RasMol [7]. They are also accessible on the Internet via a World Wide Web (WWW) browser, PDBBrowse [8], and an enhanced 3DB Browser as illustrated in Appendix C below, at Brookhaven (<http://www.pdb.bnl.gov/>) and several mirror sites worldwide.

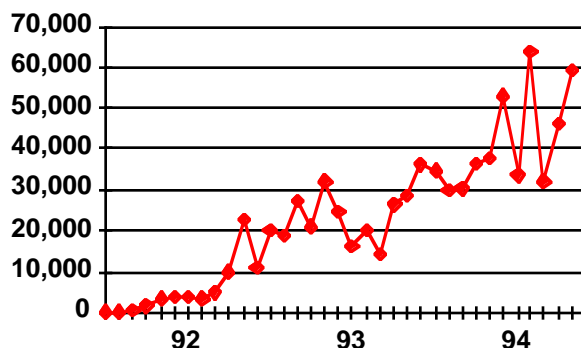


Figure 2 FTP data access

The PDBBrowse incorporates a number of features that make it easy to access information found in PDB entries. Multiple search strings covering various fields, corresponding to PDB record types such as compound, header, author, biological source, or heterogen data, are supported. These searches support boolean 'and', 'or', and 'not' operators. Entries selected can be retrieved automatically, and the molecular structures can be displayed using the public-domain X-based molecular viewer RasMol (or similar viewer). They also include hypertext links to SWISS-PROT [10], BMRB [11], the Enzyme Commission Database, and the Entrez Reference Database [13]. Internet access to the archives has become the primary mode of retrieving entries from the PDB. However, we continue to receive a considerable number of orders for our CD-ROM product. We anticipate that this will continue to be true for a variety of reasons. For example, network performance still remains poor in a number of locations and these disks, released quarterly, provide local access to the contents of the archive. Some of these network access difficulties are easily overcome by installing a copy of the PDB FTP and WWW servers using mirroring software. With this software all files in the PDB are stored locally and any changes are automatically reflected on a daily basis.

3 The 3-Dimensional Database of Biomacromolecular Structures (3DB)

Implementation of the conversion of PDB to 3DB is entailing changes in every aspect of current operations. A new data submission and archival system is being implemented, which attempts to balance the need for full automation with the need to maintain very high levels of

data accuracy and reliability. The new system relies on an RDBMS for data management and archival. An overview of the relationships between 3DBase and depositors, users, third-party software developers, and other databases is shown in Fig. 3. The following sections provide a summary of development work.

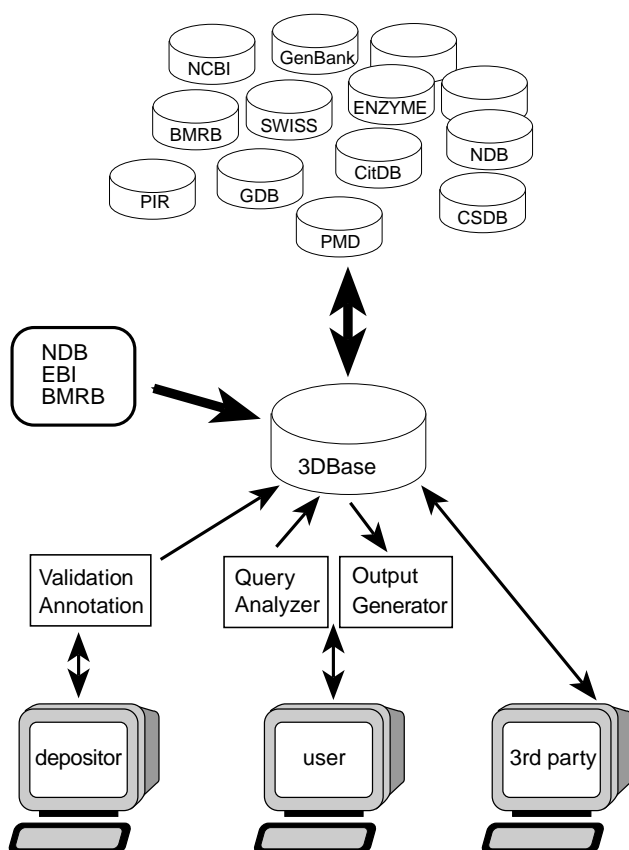


Figure 3 3DBase relationships.

3.1 The 3DBase - A Relational Database Management System for 3DB

The 3DBase is constructed with the SYBASE RDBMS, the Object-Protocol Model (OPM), and the OPM data management tools [14] developed by Victor Markowitz's group at Lawrence Berkeley National Laboratory. SYBASE provides a powerful and robust environment for data management, the OPM tools allow rapid development of SYBASE databases, and OPM's object-oriented view provides a scientifically intuitive representation of data. Along with a graphical schema editor, Markowitz's group distributes a number of other development tools. Foremost of these is a schema translator that generates SQL statements for building tables, indices, constraint rules, and triggers.

This development effort attempts to address the needs of the diverse user community served by the PDB.

The schema supports queries by those interested in answers to both crystallographic and molecular biology questions. The system is being designed with the expectation that it will shortly be federated with other biological databases. Our hope is that federation will permit complex queries to be submitted to our database, returning a composite answer built from a set of diverse databases. Interoperability is addressed through the use of schema sharing with other OPM-based databases and support for a variety of data interchange formats in the query results. In addition to providing users with a powerful environment to do complex ad-hoc queries, 3DBase will also facilitate management of the growing archive, which is expected to contain over 30,000 structural reports by the year 2000.

This work is being done as a collaboration among the following groups:

- The Protein Data Bank - Brookhaven National Laboratory
- Bioinformatics Unit - Weizmann Institute of Science
- OPM Data Management Tools Project - Lawrence Berkeley National Laboratory
- The Genome Data Base - Johns Hopkins University

3.2 Schema Development

OPM is a semantic data model that includes constructs that are powerful enough to represent the diversity and complexity of data found in PDB entries. OPM has constructs such as object class, object attribute, class hierarchy and inheritance, and derived attribute. A schema for 3DBase has been developed using OPM and is available for perusal through the PDB WWW home page. Among its notable features is a description of the coordinate data set from two perspectives. The object class oExperiment provides users with the classical view of a PDB entry which is a report of crystallographic or NMR analysis. An alternative view is presented in the class oMacroMolecule that describes the biologically active form of the molecule. Appendix A provides a description of these object classes. A clear example that demonstrates the differences between these classes is the case of the hemoglobin molecule. The oExperiment object contains the coordinates for the crystallographic asymmetric unit, which is usually a dimer. The full tetramer will, however, be presented in the oMacroMolecule object. The latter case is normally what molecular biologists are interested in when accessing PDB entries. However, crystallographers wishing to do packing studies or further refinement will need access to the oExperiment object.

In 3DBase, literature citation data are being loaded into the CitDB database of references that was developed by GDB [15]. A pointer to the appropriate entry in CitDB is loaded in the oExperiment object of 3DBase. This is an example of the strategy that we are following in linking to external databases. CitDB will be

managed as a federation run by a number of database centers that include GDB and PDB. There are several advantages to this scenario. By sharing the schema and management of the citation databases, access to information stored in each of the databases via the bibliographic citation becomes straightforward. Duplication of effort is also minimized. Today, it is still common to have several public databases build and maintain their own bibliographic databases. This will no longer be economically feasible with the expected rapid growth in database size.

3.3 Building Semantic Links to External Data Sources

Links to contents of sequence databases are provided in 3DBase via the `oPrimarySeq` and `oSeqAdv` classes. These classes form another set of objects that link 3DBase objects to external databases. Representing, building, and maintaining these links will be one of 3DB's primary tasks in the coming years. There are several issues that must be addressed for this effort to succeed. Data representation issues are foremost. Each database uses different data models to represent and store information. Semantic contents are rarely the same; for example, the primary sequence data stored in sequence databases such as SWISS-PROT or PIR [16, 17] are presented using a view which differs significantly from that used by PDB.

In general, PIR and SWISS-PROT entries contain information on the naturally occurring wild-type molecules. Each entry normally contains the sequence of one gene product and some entries include the complete precursor sequence. Annotation is provided to describe residue modifications. In both databases, the residue names used are limited to the 20 standard amino acids.

In contrast, PDB entries contain multichain molecules with sequences that may be wild type, variant, or synthetic. Sequences may also have been modified through protein engineering experiments. A number of PDB entries report structures of domains cleaved from larger molecules.

The `oPrimarySeq` object class was designed to account for these differences by providing explicit correlations between contiguous segments of sequences as given in PDB ATOM records and PIR or SWISS-PROT entries. Several cases are easily represented using this class. Molecules containing heteropolymers will be linked to different sequence database entries. In some cases, such as those PDB entries containing immunoglobulin Fab fragments, each PDB chain may be linked to several different SWISS-PROT entries.

This facility is needed, because these databases represent sequences for the various immunoglobulin domains as separate entries. `oPrimarySeq` should also be able to represent molecules engineered by altering the gene (fusing genes, altering sequences, creating chimeras, or circularly permuting sequences). In addition, it will be

possible to link segments of the structure to entries in motif databases (*e.g.*, PROSITE [18], BLOCKS [19]).

Initial building of these links is straightforward and requires analysis of a few entries coming out of a FASTA [20] or BLAST [21] search against the sequence databases. What may be problematic in the long run will be updating these links as new experimental evidence is encountered, leading to a correction in either database. Both PIR and SWISS-PROT have similar problems as they build pointers to PDB entries. To help obviate these difficulties we have agreed to establish a closer interaction between the databases. We are setting up a protocol that will broadcast to each database changes that could, in turn, affect specific entries.

4 Data Deposition

3DB will operate as a direct-deposition archive, providing mechanisms that will allow depositors to load data with minimal staff intervention. This strategy is essential if 3DB is to meet present projections of exponential growth in depositions against a fixed staff size. This is particularly challenging due to the complexity of the data being handled, the need for a common viewpoint of the entry description, and the community requirement that these data be accessible immediately upon receipt.

With direct deposition, there will be a concomitant need to increase the power of data validation procedures. These procedures must reflect current models for identifying errors and must be as complete as possible. Quality control issues assume a more central and difficult role in direct deposition strategies. Distributed data must be of the highest quality; otherwise, users will lose their trust in the archived data and will have to revalidate data received from 3DB before using them, clearly an unproductive scenario.

4.1 Current Data Deposition Procedures

Since its inception in 1971, the method followed by the PDB for entering and distributing information has paralleled the review and edit mode used by scientific journals. Currently, the author submits information which is converted into a PDB entry and is run against PDB validation programs by a PDB processor. The entry and the output of the validation suite are then evaluated by a PDB scientific staff member, who completes the annotations and returns the entry to the author for comment and approval. Table 2 summarizes checks included in our current data validation suite. Corrections from the author are incorporated into the entry, which is reanalyzed and validated before being archived and released.

Originally data flow was a manual system, designed for a staff of 1-2 scientists, and a deposition rate of about 25-50 entries per year. One person processed an entry from submission through its release. By the late

1980s, when the first steps at automation were being introduced, running the validation programs took about 4 hours per entry. Today, the same step, which includes a vastly improved set of validation programs, takes about one minute. Graphical viewing of data, a useful and powerful annotating and checking tool, has been available to processors since 1992.

Class	What is checked
stereochemistry	bond distances & angles, Ramachandran plot (dihedral angles), planarity of groups, chirality
bonded/non-bonded interactions	crystal packing, unspecified inter- and intraresidue links
crystallographic information	Matthews coefficient, Z-value, cell transformation matrices
noncrystallographic transformation	validity of noncrystallographic symmetry
primary sequence data	discrepancies with sequence databases
secondary structure	generated automatically or visually checked
heterogen groups	identification, geometry and nomenclature
miscellaneous checks	solvent molecules outside the hydration sphere, syntax checks, internal data consistency checks

Table 2 Data validation with current system

The current deposition load of ~100 entries a month is handled by about ten staff members who annotate and validate entries. The process is a production line in which checking is repeated at various steps to ensure that errors and inconsistencies in data representation are minimized. Prior to June 1994, a significant number of depositions required that administrative staff manually input information provided in a deposition form. Introduction of the current Electronic Deposition Form, together with a new parsing program, has greatly reduced hand entry of information.

Today, most of the processing time is spent resolving data representation issues and ensuring that outliers are identified and annotated. The most troublesome areas are consistently those involving handling of heterogens, resolving crystal packing issues, representing molecules with non-crystallographic symmetry, and resolving conflicts between the submitted amino acid sequence and that found in the sequence databases. Publications and other references are sometimes consulted to verify factual information such as crystal data, biological details, reference information, etc.

Processing programs, although much improved over those used in 1991, still allow errors to pass undetected through the system, requiring a visual check of all entries. We continually improve these programs and also acquire software from collaborators to address deficiencies that both we and our users have identified. In addition, we now have formed a quality control group that will be identifying sources of errors and recommending steps to improve data quality.

4.2 Development of Automatic Deposition and Validation

3DB must overcome many challenges for direct deposition to work. In a recent workshop held to assess the needs of 3DB users, crystallographers and NMR spectroscopists were unanimous in their desire for a system that did not require additional work on their part when depositing data. On the other hand, consumers (who included these same depositors) were vocal in their desire for entries to contain more information than is currently available within the PDB. We are striving to develop a suite of deposition and validation programs that accommodates these somewhat conflicting desires while ensuring that the archives maintain the highest standard of accuracy. A schematic of the automatic deposition process is depicted in Fig. 4.

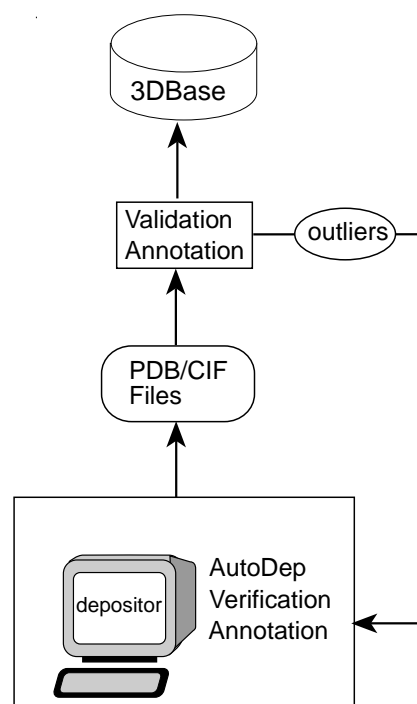


Figure 4 3DBase automatic validation.

A considerable variety of information is archived about each structure which must be supplied by the authors. The AutoDep program, currently under

development, is expected to simplify the deposition process. It includes a convenient and interactive electronic deposition protocol that guides the author in providing information. It also contains tools for data verification and validation, and is able to flag errors in syntax or spelling. The form requests approximately 500 items, including a description of the experiment and the molecule under study. Steps are being taken to help ease the burden of filling out this form. For example, the program can fill in fields using data from an existing PDB entry. These data can then be modified to reflect the contents of the new deposition. Checks against other databases are an important and evolving part of this process. Thus, names of organisms are checked against the taxonomy database of the National Center for Biotechnology Information (NCBI) [22], chemical names against IUPAC [23] nomenclature tables, and author names and citations against MEDLINE [24] (CitDB when it becomes available). FASTA/BLAST programs are run against the SWISS-PROT and PIR databases to verify protein sequences, and variant and mutant sequences are checked against the Protein Mutant Database [25] (PMD). Links between the PDB/3DB entry and these databases are established in the process. To handle the increasing number of entries with nonstandard residues (heterogens), a standard residue and heterogen dictionary is being developed to be used in the data entry and checking process. We are also adopting programs developed by the Cambridge Crystallographic Data Center [26] (CCDC), and elsewhere, to handle heterogens automatically for use in AutoDep.

In addition to the deposition form that is filled out by AutoDep, authors are requested to submit the coordinate data entry and other experimental data files for processing and archiving. Facilities are provided by AutoDep that simplify this process. An FTP script is provided that uploads author-specified local filenames to the PDB server site.

The completed form is then converted automatically into a file in PDB format and, along with the coordinate data, is submitted to a set of validation programs for checking and further annotation. These programs are designed to check 1) the quality, consistency, and completeness of the experimental data; 2) possible violations of physical or stereochemical constraints (*e.g.*, no two atoms in the same place, appropriate bond angles, etc.); 3) compliance with our data dictionary (syntax checks); and 4) in the near future, the correspondence of the experimental data to the derived structure. Development of the validation suite will evolve with advice from the community and encompass programs currently in use, written both within and outside PDB.

The validation software automatically generates, and includes in the entry, measures of data quality and consistency, as well as annotations giving details of apparent inconsistencies and outliers from normal values. This output is returned to the depositor for review. Entries whose data quality and consistency meet appropriate

standards may then be sent by the depositor directly for final review by the PDB staff and entry into the database. Entries that do not pass the quality and consistency checks may be revised by the depositor to correct inadvertent errors; alternatively, more experimental work may be needed to resolve problems uncovered.

Apparent inconsistencies or outliers may remain in a submitted entry, provided these are explained by the depositor in an annotation. In the most interesting cases, unusual features are a valid and important part of the structure. However, all such entries will be reviewed for possible errors by 3DB staff, who may discuss any important issues with the depositor. 3DB staff will then forward acceptable entries to the database.

To make automatic deposition as easy as possible, we are working with developers of software commonly used by our depositors. By modifying these programs to produce compliant data files and performing validation and consistency checks before submission, it may be possible to bypass most of the tedious steps in deposition. We are already working with Dr. A. Brünger to use procedures available through X-PLOR [27] to replace part of the validation suite for structures produced by X-ray crystallography and NMR. Diagnostic output will be included automatically as annotations in the entry. A limited version of X-PLOR will be available from BNL to all depositors for validation purposes only.

Validation of coordinate data against experimental X-ray crystallographic data requires access to structure factor data that are requested by PDB, the International Union of Crystallography (IUCr), and some journals, but are not always supplied by the depositor. We are working toward building consensus in the community that structure factor data are a necessary component of deposits of structures derived by X-ray crystallography. Statistics such as number of F's and R-values vs. $\sin(\theta)/\lambda$, etc. will be calculated and included in the 3DB entry as annotation for the experiment.

In order to make it easier for depositors to submit structure factors (as well as to exchange these data between laboratories), the PDB, in close collaboration with a number of macromolecular crystallographers, has developed a standard interchange format for these data. This standard is in CIF (Crystallographic Information File) [28, 29] and was chosen both for simplicity of design and for being clearly self-defining, *i.e.*, the file contains sufficient information for the file to be read and understood by either a program or a person. Details of this format are available through the PDB WWW server.

A consensus is still developing in the NMR community as to what types of experimental data should be deposited and what kinds of validation and consistency checks should be performed. Structural data produced by other methods may also have special features that should be archived or checked, for example the sequence alignment used for modelling studies. Requirements for the types of data to be deposited and proper ways of checking the validity and consistency of the data will be

developed in cooperation with the experimental community for each category of structure data archived by the 3DB.

4.3 AutoDep's Most Important Features

- The ability to fill in the form automatically from an existing PDB entry or from a previous deposition. When the depositor pushes a button, AutoDep will enter data from the designated file to the appropriate fields in the new form. The author merely has to update fields to reflect the new structure.
- X-ray structural refinement software is available to write PDB records that can be automatically merged into the deposition form. For example, the new releases of X-PLOR [27] and SHELX-96 write refinement details as PDB records which will be read by the program and entered in the relevant sections. We are continuing to work with authors of various programs and anticipate that increasing numbers of programs will be integrated with PDB.
- Each session has help files, examples, and links to related documentation and useful URLs to support the author during the AutoDep session.
- At any time during the AutoDep session, the Deposition Form or the resultant header portion of the PDB file can be viewed to check progress.
- The AutoDep session can be interrupted at any time and resumed hours, days, or even weeks later. The session ID number and password must be recorded to continue with the same deposition.
- When an author is satisfied with the completed Deposition Form, a submit button is provided that initiates the following:
 - The coordinates are passed through a syntax checker.
 - If they fail, the depositor is asked to correct the problem and resubmit the coordinates.
 - If they pass, the depositor is immediately sent an acknowledgement letter containing the PDB ID code.
- The entry enters the PDB processing flow.

5. Accessing Data in 3DBase

User queries to 3DBase will be via the network using general purpose graphical user interfaces such as Mosaic and Netscape. Access will also be possible through the use of software developed by third parties (commercial developers). As diagrammed in Fig. 5, user queries will be addressed to the Query Analyzer (3DB-QA), a program module running at the server site that will parse queries and pass them on to 3DBase. Query results will be

returned through the Output Generator (3DB-OG) in the format requested by the user. Queries placed over the network will generally be in the form of URLs, which are easily generated from HyperText links, HTML-based forms, or by use of programs or scripts employing the National Center for Supercomputing Applications libraries [30] for more sophisticated applications. As part of the query the user may specify the format of the response, as we do at the present time in the PDB WWW browser. The response will be frequently in the form of an HTML document, but it can also be a PDB- or CIF-formatted file [28, 29]. The conversion is done using `pdb2cif` [32]. The information returned may be either a complete or partial entry, and may include information from linked databases or external programs.

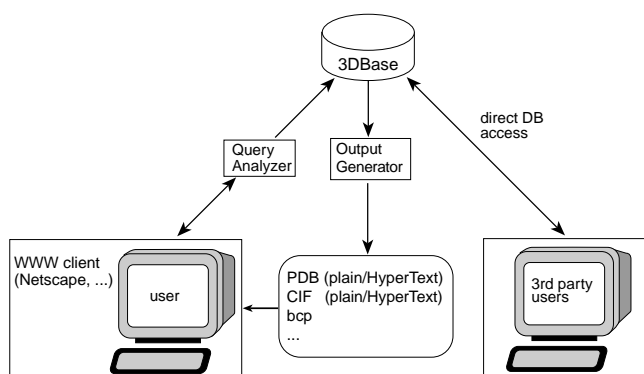


Figure 5 Accessing 3DBase

A 3DBase browser has been built using Stan Letovsky's Genera system. Users specify search criteria by filling out an HTML form. Software at BNL processes this form and generates the required SQL. System performance is improved by using stored SYBASE SQL procedures that access each predefined object. The fields available are similar to those in our PDBBrowse program, and answer most of the questions that users have been asking.

For those familiar with (or willing to learn about) the OPM protocol, access to the object layer will be provided using a high-level, OPM-based query language. As part of the 3DB open database policy, direct access to the underlying RDBMS will be allowed and actively supported. These queries are not parsed by the 3DB-QA module, so better response time can be expected. This provides third-party developers with the opportunity either to incorporate SQL clients in their products or to learn more of the OPM protocol and, thereby, gain access to all of the benefits that the Object model affords (e.g., active external links, programs, etc.). As depicted in Fig. 5, the output generator will return query results using a variety of data interchange formats. PDB will continue to support its current format in the foreseeable future. We also plan to extend this format to allow us to represent objects being stored in 3DBase. In addition, a "raw

format" is being provided which returns an attribute/value pair. This form is easily parsed and is more compact than the PDB format.

6. References

- [1] F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer, Jr., M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi, The Protein Data Bank: a Computer-based Archival File for Macromolecular Structures, *J. Mol. Biol.*, Vol. 112, pp. 535-542 (1977).
- [2] E. E. Abola, F. C. Bernstein, S. H. Bryant, T. F. Koetzle, and J. Weng, Protein Data Bank, in Crystallographic Databases - Information Content, Software Systems, Scientific Applications, F. H. Allen, G. Bergerhoff, and R. Sievers, eds., Data Commission of the International Union of Crystallography, Bonn pp. 107-132 (1987).
- [3] Crystallography, Protein Data Bank [announcement], *Nature New Biology* Vol. 233, pp. 223 (1971).
- [4] SYBASE SQL Server (Unix version 10.0) [computer program] (1994). Available Distributor: Sybase Inc., Emeryville, CA USA.
- [5] PDB-SHELL [computer program, on-line and CD-ROM]. Available anonymous FTP: ftp.pdb.bnl.gov.
- [6] Microsoft FoxPro Relational Database Management System [computer program]. Available Distributor: Microsoft Corporation, Redmond, Washington 98052-6399 USA.
- [7] R. Sayle, RasMol [computer program]. Available anonymous FTP: ftp.dcs.ed.ac.uk. Directory: /pub/rasmol. File: README [for filenames for different platforms].
- [8] D. R. Stampf, C. E. Felder, and J. L. Sussman, PDBBrowse - a Graphics Interface to the Brookhaven Protein Data Bank, *Nature*, Vol. 374, pp. 572-574 (1995).
- [9] M. C. Peitsch, D. R. Stampf, T. N. C. Wells, and J. L. Sussman, The Swiss 3D-Image Collection and Brookhaven Protein Data Bank Browser on the World-Wide Web, *TIBS*, Vol. 20, pp. 82-84 (1995).
- [10] A. Bairoch and B. Boeckmann, The SWISS-PROT Protein Sequence Data Bank: Current Status, *Nucl. Acids Res.*, Vol. 22, pp. 3578-3580 (1994). URL: <http://expasy.hcuge.ch/sprot/sprot-top.html>.
- [11] B. R. Seavey, E. A. Farr, W. M. Westler, and J. L. Markley, A relational database for sequence-specific protein NMR data, *J. Biomol. NMR* Vol. 1, pp. 217-236 (1991). URL: <http://www.bmrb.wisc.edu>
- [12] A. Bairoch, The ENZYME Data Bank, *Nucl. Acids Res.*, Vol. 22, pp. 3626-3627 (1994).
- [13] Entrez [on-line and CD-ROM]. National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD USA (producer). Available URL: <http://www3.ncbi.nlm.nih.gov/Entrez>.
- [14] I. A. Chen, and V. M. Markowitz, An overview of the Object-Protocol Model (OPM) and OPM data management tools, *Information Systems* Vol. 20 (5), 393-418 (1995). (Article and related information available at URL: http://gizmo.lbl.gov/DM_TOOLS/OPM/opm.html).
- [15] K. H. Fasman, A. J. Cuticchia, and D. T. Kingsbury, The GDB Human Genome Data Base anno 1994, *Nucl. Acids Res.*, Vol. 22, pp. 3462-3469 (1994).
- [16] K. E. Sidman, D. G. George, W. C. Barker, and L. T. Hunt, The Protein Identification Resource (PIR), *Nucl. Acids Res.*, Vol. 16, pp. 1869-1871 (1988).
- [17] D. G. George, W. C. Barker, H. W. Mewes, F. Pfeiffer, and A. Tsugita, The PIR - International Protein Sequence Database, *Nucl. Acids Res.*, Vol. 22, pp. 3569-3573 (1994).
- [18] A. Bairoch and P. Bucher, PROSITE: Recent Developments, *Nucl. Acids Res.*, Vol. 22, pp. 3583-3589 (1994). URL: <http://expasy.hcuge.ch/>
- [19] S. Henikoff and J. G. Henikoff, Protein family classification based on searching a database of blocks, *Genomics* Vol. 19, pp. 97-107 (1994). URL: <http://blocks.fhcrc.org/>
- [20] W. R. Pearson and D. J. Lipman, Improved tools for biological sequence comparison, *Proc. Natl. Acad. Sci. USA* Vol 85, pp. 2444-2448 (1988).
- [21] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, Basic local alignment search tool, *J. Mol. Biol.* Vol. 215, pp. 403-410 (1990).
- [22] National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD USA (producer). Available URL: <http://www.ncbi.nih.gov>. Available anonymous FTP: ncbi.nlm.nih.gov. Directory: /repository/taxonomies.
- [23] Biochemical Nomenclature and Related Documents, a compendium, second edition, International Union of Biochemistry and Molecular Biology, prepared by Claude Liebecq, Portland Press Ltd., London (1992).
- [24] MEDLINE [on-line and CD-ROM]. National Library of Medicine, National Institutes of Health, Bethesda, MD USA (producer). Available: NLM, DIALOG, BRS, SilverPlatter.
- [25] K. Nishikawa, S. Ishino, H. Takenaka, N. Norioka, T. Hirai, T. Yao, and Y. Seto, Constructing a protein mutant database, *Protein Eng.* Vol. 7, pp. 733 (1994).

- [26] F. H. Allen, J. E. Davies, J. J. Galloy, O. Johnson, O. Kennard, C. F. Macrae, E. M. Mitchell, G. F. Mitchell, J. M. Smith, and D. G. Watson, The development of versions 3 and 4 of the Cambridge Structural Database System, *Journal of Chemical Information and Computer Science* Vol. 31, pp. 187-204 (1991).
- [27] A. T. Brünger, X-PLOR - Version 3.1, A System for X-ray Crystallography and NMR, Yale University Press, New Haven (1992).
- [28] S. R. Hall, F. H. Allen, and I. D. Brown, The Crystallographic Information File (CIF): a new standard archive file for crystallography, *Acta Cryst.* Vol A47, pp. 655-685 (1991). (Related information available at URL: <http://www.iucr.ac.uk/cif/home.html>)
- [29] P. M. D. Fitzgerald, H. M. Berman, P. E. Bourne, and K. Watenpaugh, The macromolecular CIF dictionary, American Crystallographic Association Annual Meeting, Albuquerque, NM USA (1993).
- [30] National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, Illinois USA. Available URL: <http://www.ncsa.uiuc.edu>.
- [31] S. I. Letovsky, Genera [computer program] (1994). URL: <http://gdbdoc.gdb.org/letovsky/genera>.
- [32] H.J. Bernstein, F.C. Bernstein, and P.E. Bourne (in preparation) CIF Applications. pdb2cif: Translating PDB Entries into mmCIF format. Available at URL: <http://ndbserver.rutgers.edu/mmcif/software/pdb2cif>

7. Appendix A: Description of primary database objects

The following describe two primary object classes found in 3DBase. For a more complete and detailed view, you may use our schema browser at URL <http://pdb.pdb.bnl.gov/opmbrowser.html>. An OPM schema consists of definitions of object classes, each described by a set of attributes. Data types assigned to attributes can be primitive types such as numbers or character strings, or they can be other classes defined in the schema. In addition to object classes, OPM provides controlled value classes that restrict values possible for objects in the class. Attributes can be either single or multiple valued. The latter can be specified as an ordered list (list-of) or as an unordered list (set-of). Object classes are grouped into a hierarchy of subclasses and superclass relationships called an ISA hierarchy. A class is said to inherit all the attributes of its superclass. In the description below object classes have been assigned names which start with the small letter "o", attributes with the

small letter "a", and controlled value classes with the prefix "cv".

oExperiment
isa o3DBExportObj

*aTitle: list-of varchar(255)

required

Description: Contains a title for the experiment or analysis that is represented in the object. It should identify an entry in the PDB in the same way that a title identifies a paper.

*aDepositor: list-of oPerson

required

Description: An ordered list of names with address information.

*aKeywords: set-of varchar(80)

optional

Description: Set of keywords relevant to the entry. These provide a simple means of categorizing the experiment or the molecules studied.

*aExpdta: cvExperimentTypes

required

Description: Identifies the experimental technique used in the study. This normally refers to the type of radiation and sample, but can also include the spectroscopic or modeling technique. Permitted values include:

ELECTRON DIFFRACTION
FIBER DIFFRACTION
FLUORESCENCE TRANSFER
NEUTRON DIFFRACTION
NMR
THEORETICAL MODEL
X-RAY DIFFRACTION

*aReference: list-of oExternalReference

optional

Description: Publications related to the study. These citations are chosen by the depositor.

*aRemarks: set-of oAnnotations

optional

Description: General comments regarding the experiment or the molecules studied.

oMacroMolecule
isa o3DBExportObj

*a3DBID: char(4)

required

Description: Contains the PDB identification code. This is a four character field of which the first character must be an integer greater than zero. This identifier is unique within PDB and is assigned randomly to an entry.

*aMolName: set-of varchar(80)

optional

Description: Set of molecule names. Each molecule may be assigned more than one name allowing for the use of synonyms and aliases.

*aSrcDescr: oSource

optional

Description: Specifies the biological and/or chemical source of each biological molecule in the entry. Sources are described by both the common name and scientific names, genus and species. Strain and/or cell-line for immortalized cells are given when they help in uniquely identifying the biological entity studied.

*aMolType: cvMacroMoleculeType

optional

Description: Molecule type -Valid values are:

protein

DNA

RNA

polysaccharide

other - must annotate

*aBioMol: varchar(255)

optional

Description: Information on accessing the structure of the complete biological molecule. Currently this contains the filename for a biomol entry found in the PDB ftp server. This attribute will be replaced by a new object class with attributes that provide the transformation matrices, descriptive text, and, if available, the filename for the coordinate set.

*aCoordinates: set-of oChain

optional

Description: Atomic coordinate values stored as individual chains.

*aMolSequence: list-of oPrimarySeq

optional

Description: SEQRES records contain the amino or nucleic acid sequence of residues in each chain of the macromolecule.

*aDomain: varchar(100)

optional

Description: Specifies a domain or region of the molecule.

*aEngineered: cvFlagDict

optional

*aEnzyme: set-of oExternalReference

optional

Description: The Enzyme Commission number associated with the molecule.

*aMutation: varchar(255)

optional

Describes the mutations present.

*aFormula: varchar(80)

optional

*aMolWeight: float

optional

*aMolID: cvLocalID

required

Description: Integer to uniquely identify each instance of a coordinate set for a molecule. For example, each occurrence of lysozyme in the database will be identified by a unique number.

*aAnnotate (aSummLine, aExtDB) :

set-of (varchar(255), oExternalReference)

optional

Description: Annotations describing the molecule. This is presented as a table of text and pointer to an external database.

8. Appendix B: 3DBase Report in Different Formats

3DBase - raw format

For the oMacromolecule object of the entry 1ACE:

*Export_Object: 1ACE

*Macromol_name: Acetylcholinesterase

*Macromol_name: Ache

*EC_number: 3.1.1.7

*3DB_init_res_num: 4

*3DB_term_res_num: 534

*Init_res_num: 25

*Term_res_num: 555

*Database_ID_code: ACES_TORCA

*Domain_desc: No

*Engineered: No

*Source_sci_name: Torpedo californica

*Source_common_name: Pacific electric eel

Data in BoulderIO format:

Export_Object = 1ACE

Macromol_name = Acetylcholinesterase

Macromol_name = Ache

EC_number = 3.1.1.7

Chain = {

3DB_init_res_num = 4

3DB_term_res_num = 534

Init_res_num = 25

Term_res_num = 555

Database_ID_code = ACES_TORCA

}

Domain_desc = No

Engineered = No

Source_sci_name = Torpedo californica

Source_common_name = Pacific electric eel

9. Appendix C: Guided Tour to 3DB Browser

- Search for "acetylcholinesterase" where Kinemage annotation is available
- Search for "epidermal" into NMR experimental data

10. Appendix D: Guided Tour to AutoDep

10.1 Introduction

This is a description of how to submit data to the PDB using AutoDep. If you have made submissions to the PDB before, then the WWW AutoDep form is a lot like the earlier Electronic Deposition Form with the following additional features:

- You can start with an existing PDB entry or a previously completed Deposition Form.
- Your deposition is partially verified as you complete it, saving a lot of time: we catch errors while you are still available to correct them.
- We provide a script that downloads all of the required data to the PDB (Unix only).
- As you are already using the Web, hot links are provided that may be useful in constructing your submission.

10.2 Overview of the Process

1) Select a location for your submission. We are pleased to have global partners in Israel and the United Kingdom that will soon be serving as deposition sites. Sometimes, the speed of network links would dictate the use of one site over another.

2) The next step concerns security and confidentiality. We provide you with a deposition number consisting of three letters, a dash, and a number, and ask you to provide a password. Anyone who knows both the deposition code and the password has access to your deposition. This allows you to make future submissions based on earlier ones. If you prefer us to block all access at the completion of the deposition process, please instruct us accordingly. This may be done in the section entitled "Special Instructions" to the PDB. As a security measure, if a specific amount of time has passed between transactions, or the request comes from a different computer, you will be asked to re-enter the password. The number of times you should have to do this will be minimal, but it is done to protect against an unmonitored workstation being used to view your work.

3) The next step is to transfer to the PDB AutoDep server various data files that are needed to initiate the AutoDep process. This is done by:

- Selecting and filling in the section "Files Being Deposited".
- Among the files you may need to transfer are: PDB-formatted output data from an X-ray refinement program such as X-PLOR, SHELX, etc., or a PDB file containing your coordinates; topology, parameter, restraints, sequence and alignment files, etc.
- You will be given a Unix script which you may copy and paste into a shell or X window with your mouse: run this script to transfer the files. Please note that a bug exists in the SGI operating system that precludes you from using this script in a WINTERM window running tcsh. In that case, you will have to open an xterm window or run csh in one of the standard windows.
- Reload the Main Menu Page and click on the merge button to enter the data into the Deposition Form, overwriting anything already existing in those fields.

4) Next, you will be asked to complete the remaining sections.

5) You may preview the PDB file at any time during the process in order to view how it is shaping up.

6) You may also view the dep form and save a completed copy of it for your records.

7) Within seconds of your submission, you will be sent an e-mail acknowledgement that your submission has been received at the PDB.

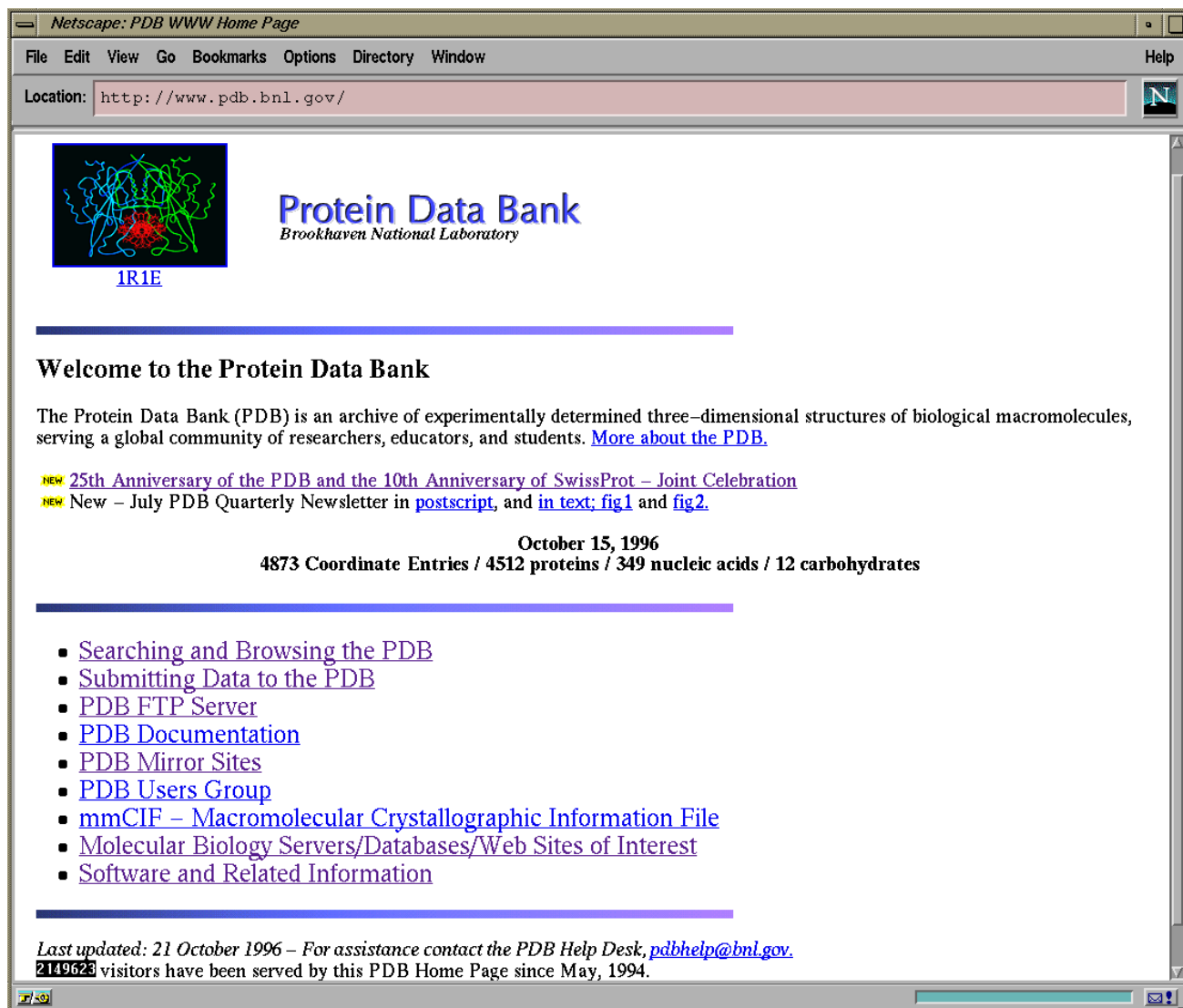
8) If the coordinates pass the syntax check you will receive your PDB ID code within seconds. If not you will be asked to make corrections to the format and resubmit your deposition.

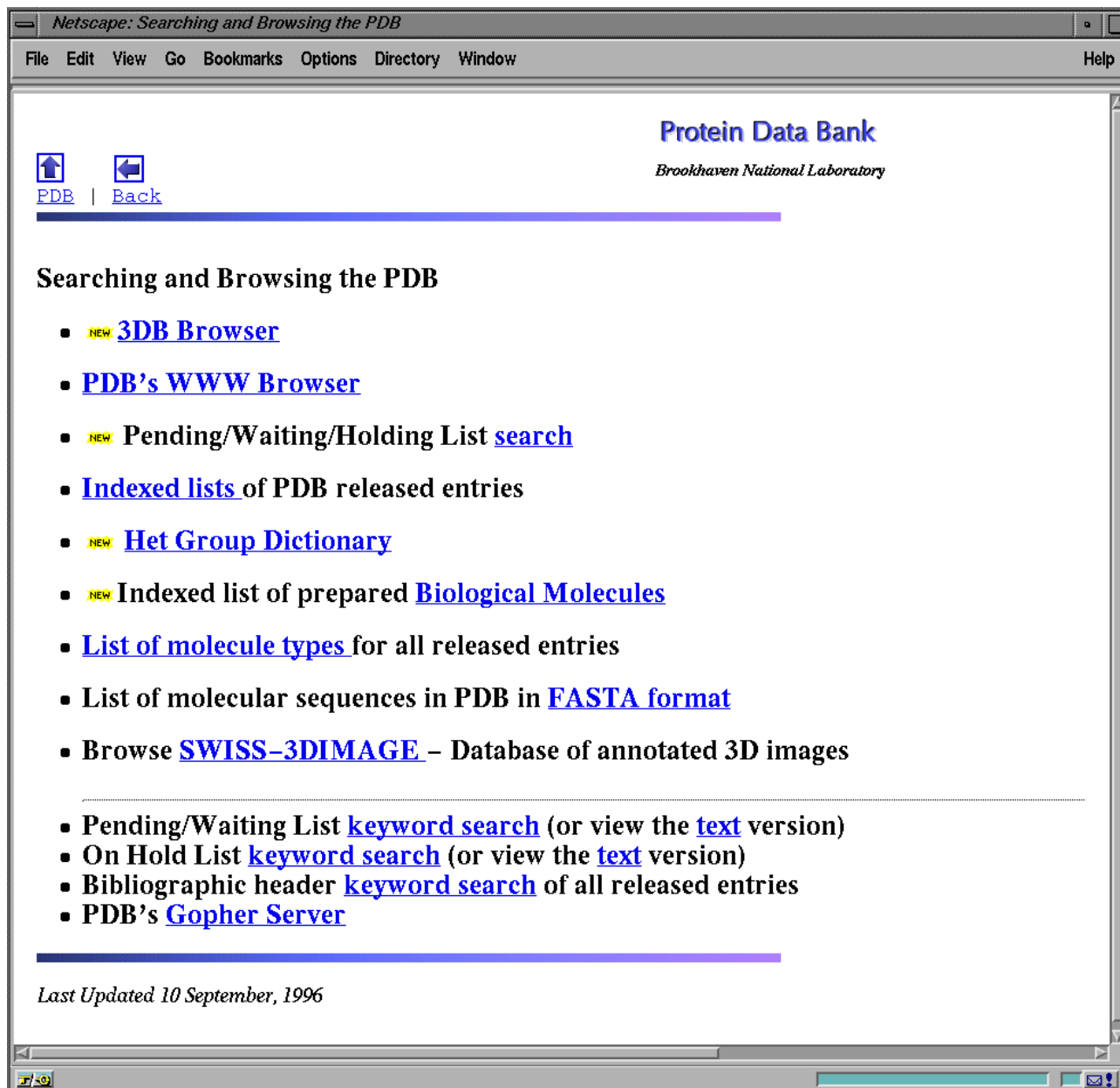
So - Let's Begin!

To which location do you prefer to submit your data?

- Brookhaven National Laboratory, New York State, USA
- Weizmann Institute of Science, Rehovot, Israel
- EMBL Outstation, European Bioinformatics Institute, Hinxton, United Kingdom

Figures for Appendix C: Guided Tour to 3DB Browser





Netscape: 3DB Browser

File Edit View Go Bookmarks Options Directory Window Help

Protein Data Bank

Brookhaven National Laboratory

3DB Browser

[READ\[Rasmol mime type\]](#)

Start a search Clear or [Upload](#) from previous query

PDB ID Enter 1a for all entries starting with '1A'.

Full text query of PDB data [☐ Exact word match]

Enter here one or more words to search for in the complete PDB entry. Try 'hemoglobin AND deoxy' or insert [AU] before the name of an author, like in '[AU] Wodak' when searching for depositors.

Additional constraints for refining your query, tailoring the results. (You may use this selection together with the Full Text Query and/or PDB ID queries)

<p>Biological unit</p> <p>Kinematic</p> <p>NMR experimental data</p> <p>Rasmol script</p>	<p>Representative structure</p> <p>Structure factors</p> <p>Text comment</p>	<p>◆ AND ◆ OR these constraints</p> <p>You may select one or more of these constraints, stating at the same time if all of them should be present (AND) or at least one of them (OR) in the resulting list of PDB ID codes.</p> <p>Start a search Clear</p>
--	--	---

Resolution examples: Enter 2.17-2.20 for a range search. Enter 3.0 for a unique value

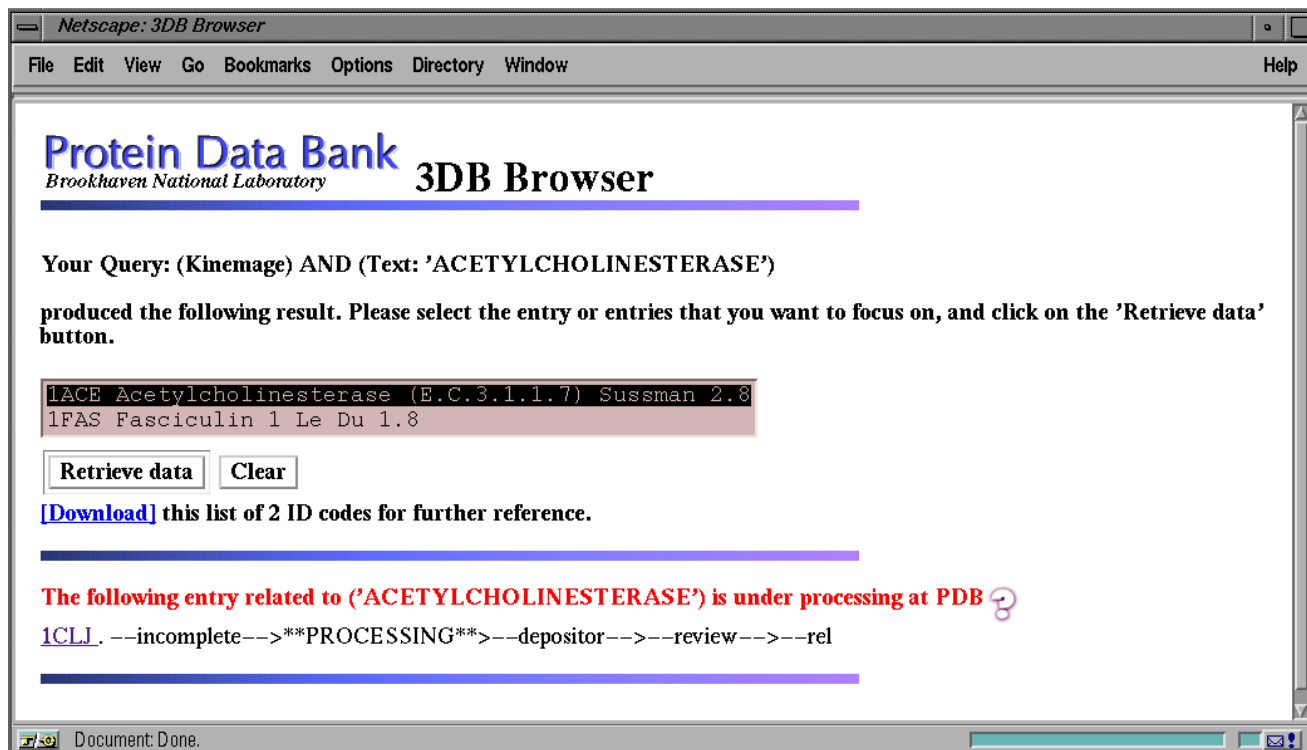
Released between (dd-mm-yyyy) and (dd-mm-yyyy)

Select only those entries released (or updated) between selected dates. Use either '/' or '-' as separators. The month can be entered as a 3 letters name (as in 9/Sep/1986) or a number (as in 31-11-1990)

Start a search Clear

Send comments and suggestions to Jaime Prilusky, lsprilus@weizmann.weizmann.ac.il

Document: Done.



Netscape: 3DB Atlas for 1ACE

File Edit View Go Bookmarks Options Directory Window Help



This is 1ACE

HEADER	HYDROLASE (CARBOXYLIC ESTERASE)	08-OCT-91	1ACE	1ACE	2
COMPND	ACETYLCHOLINESTERASE (E.C.3.1.1.7)			1ACE	3
SOURCE	ELECTRIC RAY (TORPEDO \$CALIFORNICA)			1ACE	4
AUTHOR	J.L.SUSSMAN,M.HAREL,I.SILMAN			1ACE	5
REVDAT	1 15-JAN-92 1ACE 0			1ACE	6

Data retrieval:
 Asymmetric unit, PDB entry: [\[header only\]](#) or [\[complete with coordinates\]](#)
 Retrieve 1ACE in [mmCIF](#) format

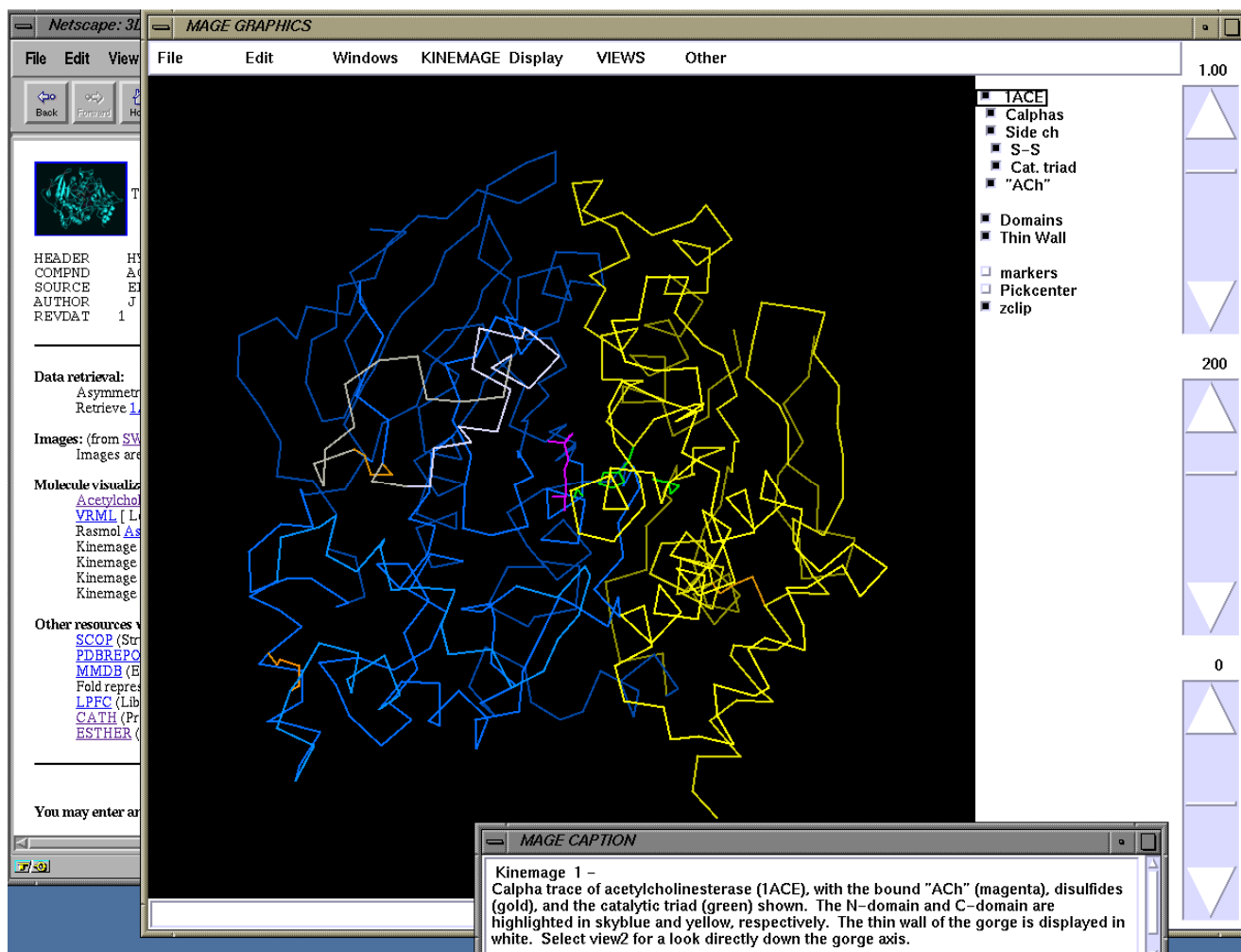
Images: (from [SWISS-3DIMAGE](#) image database at [ExPASy](#))
 Images are available in [GIF](#), [SGI](#) and [JPEG](#) format.

Molecule visualization:
[Acetylcholinesterase](#) (Nature's Vacuum Cleaner)
[VRML](#) [Look here for Virtual Reality Modeling Language [Browsers](#)]
 Rasmol [Asymmetric unit](#)
 Kinemage [Axelsen.kin](#), Active site gorge, xray & dynamics
 Kinemage [Cygler.kin](#), Family of esterases, lipases compared
 Kinemage [Mortensn.kin](#), Ser CPases, stabilizing Glu-Glu bridges
 Kinemage [Vandnbn.kin](#), Docking of the peptide fasciculin to ACE, model

Other resources with information on 1ACE:
[scop](#) (Structural Classification of Proteins)
[PDBREPORT](#) (protein verification by [WHAT_CHECK](#) procedures)
[MMDB](#) (Entrez's Structure Database)
 Fold representative is [1ack](#) from [Dali/FSSP](#) (Families of Structurally Similar Proteins)
[LPFC](#) (Library of Protein Family Cores)
[CATH](#) (Protein Structure Classification)
[ESTHER](#) (ESTerases and alpha/beta Hydrolase Enzymes and Relatives)

You may enter another PDB ID code

Document: Done.



Netscape: 3DB Browser

FileEditViewGoBookmarksOptionsDirectoryWindowHelp

Protein Data Bank

Brookhaven National Laboratory

3DB Browser

Start a search

Clear

 or [\[Upload\]](#) from previous query

PDB ID

Enter 1a for all entries starting with '1A'.

Full text query of PDB data

☐ Exact word match

EPIDERMAL

Enter here one or more words to search for in the complete PDB entry. Try 'hemoglobin AND deoxy' or insert [AU] before the name of an author, like in '[AU] Wodak' when searching for depositors.

Additional constraints for refining your query, tailoring the results.

(You may use this selection together with the Full Text Query and/or PDB ID queries)

Biological unit

Kinemage

NMR experimental data

Rasmol script

Representative structure

Structure factors

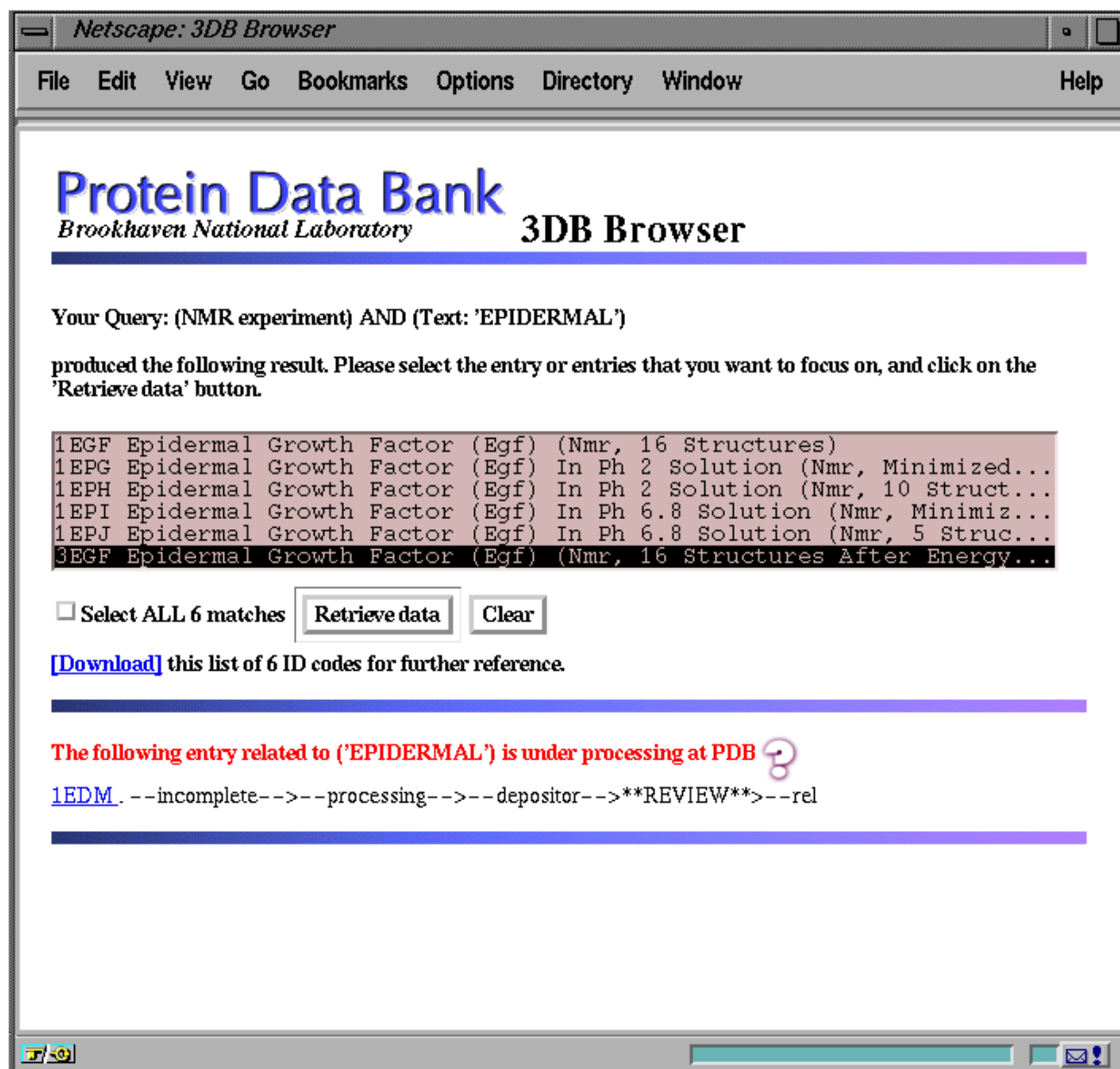
Text comment

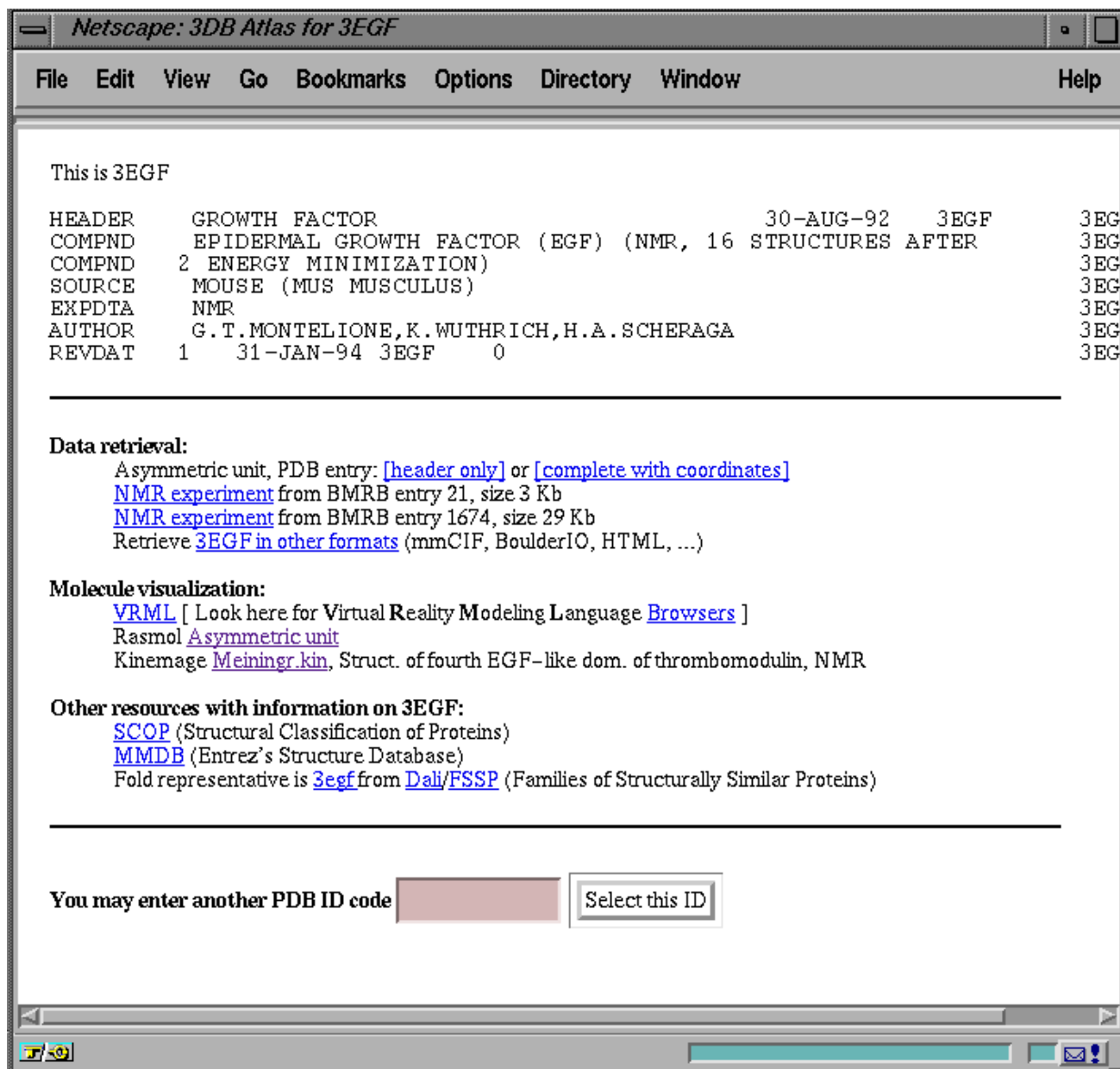
◆ AND ◆ OR these constraints

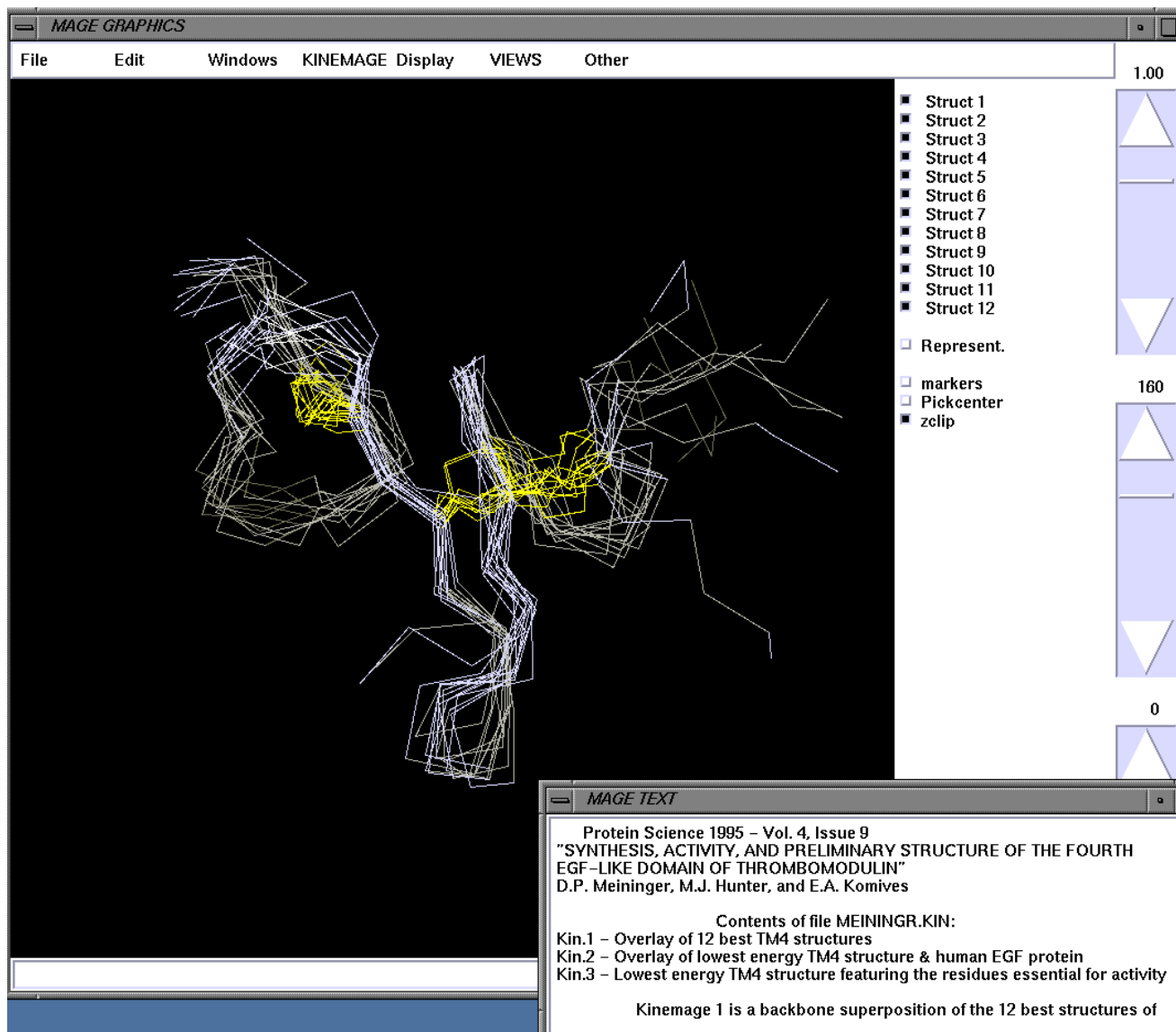
You may select one or more of these constraints, stating at the same time if all of them should be present (AND) or at least one of them (OR) in the resulting list of PDB ID codes.

Start a search

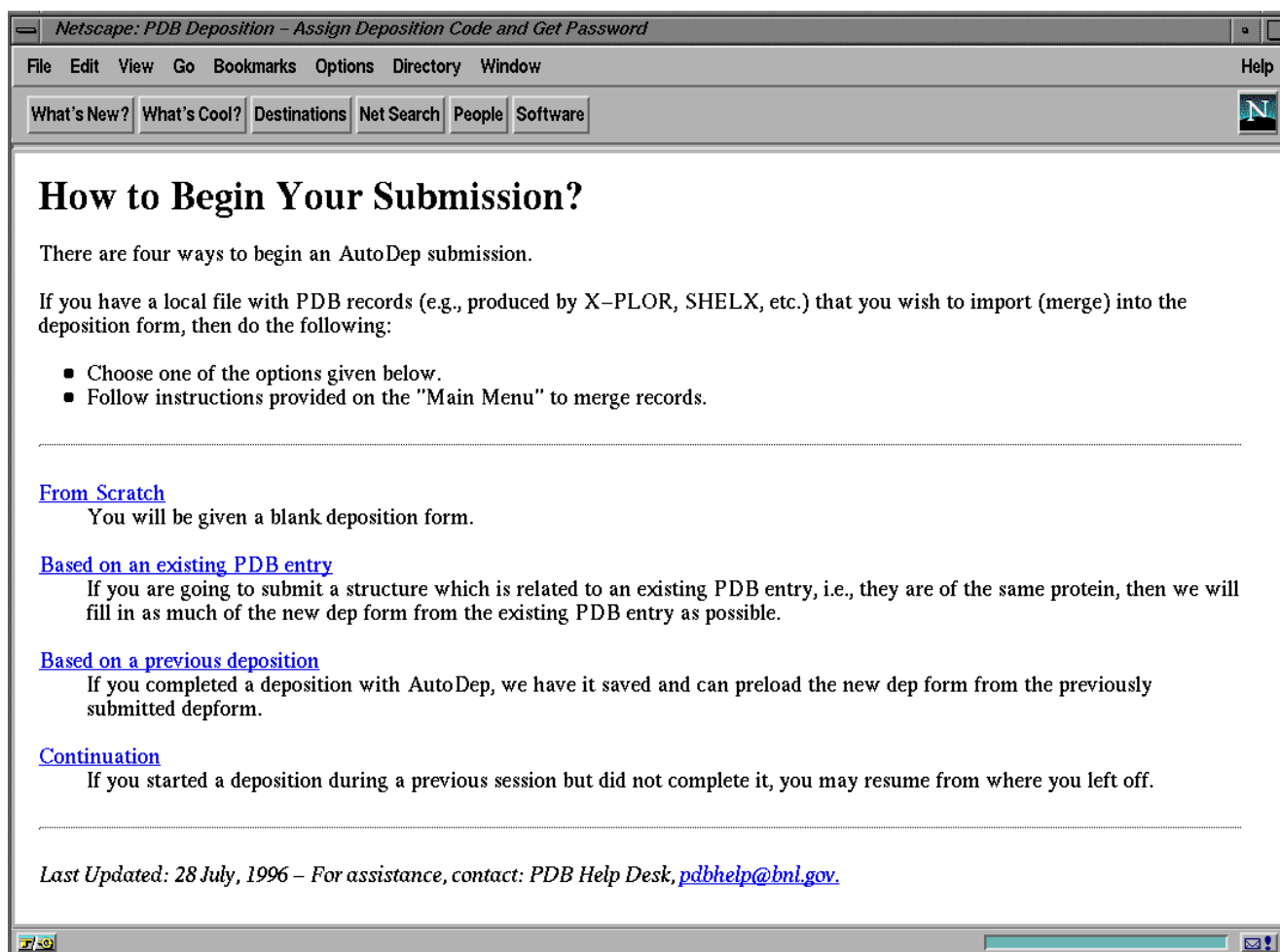
Clear







Figures for Appendix D: Guided Tour to AutoDep





Netscape: PDB Deposition – Assign Deposition Code and Get Password – New Deposition

File Edit View Go Bookmarks Options Directory Window Help

Location:

Start a New Deposition (BNL-1701)

In addition to remembering that your deposition id is **BNL-1701**, you also have to provide a password, typing it twice in order to verify the spelling. As you go through the deposition process, you may be asked to re-enter the password again. This will only happen if there is a time lapse in submitting various forms. *This is to protect the confidentiality of your entry, and we ask your understanding.*

 **DO NOT FORGET BNL-1701** 

Please input a password:

Please reenter it:

Last Updated: 28 July, 1996 – For assistance, contact: PDB Help Desk, pdhelp@bnl.gov.

Netscape: Sections of PDB Deposition Form BNL-1701

File Edit View Go Bookmarks Options Directory Window Help

Location: <http://terminator.pdb.bnl.gov:4148/cgi-bin/xdepla.pl>

Protein Data Bank

Brookhaven National Laboratory

Main Menu for deposition form BNL-1701

This page guides you to all sections of the deposition form and submits your deposition to PDB when complete. The **X** will change into a **✓** on each section below once that section is correctly filled out and verified. See the [Concordance](#) for a table of which section corresponds to which PDB records.

Click on a title to move to that section. For your entry to be accepted by the PDB, every section must have a **✓**. When you finish working on a section, return to this page by clicking on the [\[Main Menu\]](#) label that appears at the bottom of every section.

AutoDep Version 1.0 [Release Notes](#)

Deposition Sections

X Depositor Information	X NCS and Description of Biomolecule
X Files Being Deposited	X Connectivity
X Special Instructions to the PDB	X Sequence Information
X Experimental Details	X Refinement Information
X Title, Authors, and Keywords	X Heterogens
X Compound and Source	X Other Annotations
X References	X Secondary Structure
X Crystal and Coordinate System	

Merge Imported Files into Deposition Form

If you have submitted files to the PDB and those files contain records in valid PDB format, then you may merge the contents of those files with the deposition you are currently working on.

Please note that when you merge a file, you completely overwrite the fields in the deposition form that correspond to the fields in the file. There is no way to selectively merge data items. Therefore, file upload and merging should be the first thing you do when starting a new submission.

If you just uploaded files, then you must *Reload*(Netscape Menu Bar) this page in order for the file names to appear.

Netscape: Files Being Deposited BNL-1701

File Edit View Go Bookmarks Options Directory Window Help

Location: <http://terminator.pdb.bnl.gov:4148/cgi-bin/xsection.pl?depid=BNL-1701&scrambled=pdqgx2vQUyEhU&Asection=fi>


Go to the [\[Main Menu\]](#) or [\[Preview\]](#) your deposition BNL-1701

✓ Files Being Deposited BNL-1701

List the names of all files being deposited at this time for this structure. Structure factor files must be deposited with x-ray diffraction studies. Every field requires an answer, even if it is only a n/a.

References for the symbols and links to other resources and the Main Menu page can be found at the bottom of the form.

List files using their full pathname so that we can provide you with the script to submit them via FTP. After transferring the files to PDB, you must click on your Web Browser's Reload button to refresh the Main Menu page; then you may use the Merge button to write any valid PDB records within your files to this deposition form.

 (After saving you may safely return to the Main Menu.)

List each file using its complete pathname. Include identifying information within each file: your name and e-mail address, your AutoDep ID number, the file name, and the title of this PDB entry.

✓file containing the atomic coordinates,
may also contain other PDB records (e.g., /usr/sam/xyz.pdb)

✓file(s) containing PDB header records,
if not in the atomic coordinate
file (e.g., /usr/sam/xplor_output.pdb)

✓structure factors file (e.g., /usr/sam/x.sf)

✓format of the SF file (e.g., X-PLOR format)

✓NMR restraints file (e.g., /usr/sam/x.nmr)

