July 2002 • Release #101

[redacted]

The Protein Data Bank is operated by Rutgers, The State University of New Jersey; the San Diego Supercomputer Center at the University of California, San Diego; and the National Institute of Standards and Technology — three members of the Research Collaboratory for Structural Bioinformatics (RCSB).

Protein Data Bank Advisory Notice

The PDB is supported by funds from the National Science Foundation, The Office of Biological 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. It is a common goal of these agencies and the managing institutions of the PDB to make public and disseminate as widely as possible information compiled and archived by the PDB.

By using the materials available in this archive the user agrees to abide by the following conditions.

· The archival data files managed by the PDB archive are made freely available to all users. Data files within the archive may be redistributed in original form without restriction. Any redistribution of all or a subset of entries that have been modified yet retain their original PDB identifier(s) (e.g. 4HHB) is prohibited.

· Use of the PDB or data contained within this resource should be acknowledged using the following citation: H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne: The Protein Data Bank. Nucleic Acids Research, 28 pp. 235-242 (2000).

· The user assumes all responsibility for insuring that intellectual property claims associated with any data set deposited in the PDB are honored. It should be understood that the PDB data files do not contain any information on intellectual property claims with the exception in some cases of a reference for a patent involving the structure.

Liability Disclaimer

The resources of the PDB are provided on an "as is" basis. The institutions and individuals managing the PDB cannot be held liable to any party for direct, indirect, special, incidental, or consequential damages, including lost profits, arising from the use of the archived materials.

The resources provided by the PDB are provided WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESSED OR IMPLIED. THIS INCLUDES BUT IS NOT LIMITED TO MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. THE INSTITUTIONS MANAGING THE PDB MAKE NO REPRESENTATION THAT PDB RESOURCES WILL NOT INFRINGE ANY PATENT OR OTHER PROPRIETARY RIGHT.

Any opinion, findings and conclusions or recommendations expressed in the PDB by the authors/contributors do not necessarily reflect the views of the government agencies listed above, or the managing institutions of the PDB.

The Protein Data Bank CD-ROM Set

July 2002

The Protein Data Bank (PDB) is an international repository for macromolecular structure data, generated experimentally by X-ray crystallographic and NMR methods, or from theoretical modeling and other techniques. This CD-ROM set contains the macromolecular structure data released through July 1, 2002.

Table of Contents

I. Changes in the CD-ROM Set

18,528 Structures are included in this nine CD-ROM set.

Changes to be implemented with the October 2002 PDB CD-ROM release

1. Theoretical structure models will be on disk 1 in directory *models*.

2. Beginning with the October 2002 CD-ROM release, subscribers will only receive the coordinate files for macromolecular structure entries on the quarterly CD-ROM sets.

Users wishing to receive experimental data files on CD-ROM should sign up using the form in the *Readme.doc* file on disk 1, directory *pub* or using the web order form accessible from *http://www.rcsb.org/pdb/cdrom.html*.

All CD-ROM services will continue to be available free of cost.

Changes implemented with this release

None

For a list of previous changes please consult the *Readme.doc* file on disk 1, directory *pub*.

Your comments and suggestions are welcome. Please send them to ***info@rcsb.org*** or fax to +1 301 975 8717 or mail to PDB, Mail Stop 8314, Gaithersburg, MD, 20899-8314.

II. What is on the CD-ROM set

The Protein Data Bank as of July 1, 2002 is contained on this CD-ROM set. The total number of structures in this set is 18,528.

A. Coordinate, Structure Factor, and NMR Constraint Data

The purpose of this CD-ROM is to make available, for research and instruction, the coordinates of X-ray or NMR determined macromolecular structures. When available, the X-ray structure factor or NMR constraint files used in the structure determination are also included.

A few files in the archive have headers but no coordinates. Those files are not compressed and have an extension of *.noc*.

1. Directory Structure

The coordinate files are found on each disk in a directory named *entries*. Structure factors files are found in a directory named *strucfac*. NMR constraint files are in a directory named *nmr_mr*. Coordinate, structure factor and NMR constraint files are managed in a series of subdirectories that are named so that the directory name is the middle two characters of the PDB ID codes in that directory. For example coordinates file *1dom* would be found in directory *do*. The structure factor data that are not in the mmCIF format are an exception and are found in a subdirectory named *nonCIF*. The coordinates, structure factors and NMR constraints are divided across the CD-ROM set in *entries*, *strucfac*, and *nmr_mr* directories as follows:

Disk 1 - two letter directories 00-ba

Disk 2 - two letter directories bb-d4

Disk 3 - two letter directories d5-el

Disk 4 - two letter directories em-g7

Disk 5 - two letter directories g8-hx

Disk 6 - two letter directories hy-jx

Disk 7 - two letter directories jy-mr

Disk 8 - two letter directories ms-r0

Disk 9 - two letter directories r1-zz

2. Compression

The files have been compressed using the gzip software. Copies of that software (for some UNIX, PC and Mac systems) that can be used to uncompress the files are on DISK 1 in directory */pub/xtrnl_sw*.

3. ISO 9660 Constraints

The CD-ROMs are written in standard ISO 9660 format. ISO 9660 format limits file names to eight characters and only allows for one file extension. Consequently after the files were compressed their names were changed to omit the characteristic *.ent* or *.mr* extension and leave only the *.gz* extension so that they could be written to the CD-ROM.

After the files have been transferred to your disk drive and uncompressed, it is possible to add the extension back. The following code for UNIX systems was suggested on earlier copies of the PDB CD-ROM. This example is for the coordinates in directory *entries* that need to have the *.ent* extension.

```
foreach dir (/entries/*)
        foreach lfn ($dir/*)
                mv $lfn $lfn.ent
        end
end
```

B. Other Resources

In addition to the atomic coordinates, NMR constraints, and X-ray structure factors, Disk 1 contains a directory named *pub*. That directory contains an assortment of files and directories as follows:

1. Indices derived from the PDB files

The CD-ROM set contains files relating specific fields in the files to the PDB ID code. The layout of these files varies; some are delimited some are not. To use them refer to the specific file. They are found on disk1 in /pub/resource/index. These files can also be found on the PDB ftp site at ftp.rcsb.org/pub/pdb/derived_data/index

author.idx – idcode, author name

cmpd_res.idx – idcode, resolution, compound name as found in the compound record. The resolution value for entries derived from NMR or theoretical model studies is -1.00. This file is ordered by compound name.

compound.idx – idcode and compound name as found in the compound record.

crystal.idx – idcode, unit cell parameters, space group, Z value.

entries.idx – idcode, header information, accession date, compound, source, author list, resolution, experiment type (sometimes omitted for X-ray).

obsolete.dat – date, obsolete PDB ID, successor PDB ID where appropriate. This file is the same as the file by this name in */pub.*

resolu.idx – idcode, resolution. The resolution value for entries derived from NMR or theoretical model studies is -1.00.

source.idx – idcode, source name as found in the compound records. This file is ordered by source.

2. Deposition Forms

Deposition forms for both X-ray and NMR depositions are found in */pub/dep_nmr.txt* and */pub/dep_xray.txt*. They can also be found online at *ftp.rcsb.org/pub/pdb/doc*. These mmCIF forms can be used to submit NMR or X-ray structures to the PDB as an alternative to the Web-based deposition tool, ADIT (**http://deposit.pdb.org/**).

3. Other information documents

These documents are listed by content - file name in italics - and records or subject matter. On-line availability for each file is given.

a) Holdings – */pub/holdings.doc* or */pub/holdings.htm* – A breakdown of the numbers of PDB structures by experimental technique and molecule type. More extensive PDB growth and holdings information, including graphical depictions, are available on-line at: ***http://www.pdb.org/pdb/holdings.html.***

b) Entry types – */pub/entrytyp.txt* – idcode, type as protein or nucleic acid and experimental technique as *diffraction* for X-ray diffraction, *NMR*, or *model*. That file can be found on-line at: ***ftp://ftp.rcsb.org/pub/pdb/derived_data/.***

c) PDB Contents Guide – */pub/cntnt_21.txt* – Atomic Coordinate Entry Format Description, Version 2.1. A guide to the 1996 version of the PDB file format gives a complete description of the contents of PDB coordinate entry files. Plain text as well as html copies can be found online at:
http://www.rcsb.org/pdb/docs/format/pdbguide2.2/guide2.2_frame.html,

ftp://ftp.rcsb.org/pub/pdb/doc/format_descriptions/Contents_Guide_21.txt,

ftp://ftp.rcsb.org/pub/pdb/doc/format_descriptions/Contents_Guide_21.html.

d) Het dictionary – */pub/hetgroup*/*het_dict.txt* – picture, residue, het synonym, het name, and formula. This file can also be found at: ***ftp://ftp.rcsb.org/pub/pdb/data/monomers/het_dictionary.txt.***

e) PDB Newsletter – *nwslettr.doc* – The newsletter is available on the CD-ROM set on disk 1, in the directory *pub*, as a text file. All the PDB Newsletters are available on-line at
http://www.pdb.org/pdb/newsletter.html, or
ftp://ftp.rcsb.org/pub/pdb/doc/newsletters.

f) Newsletter subscription information – */pub/nwsl_sub.doc* – email address to sign up for the PDB Newsletter. To subscribe online please go to ***http://www.rcsb.org/pdb/forum.html***.

g) Obsolete records – */pub/obsolete.dat* – date, idcode, and replacement idcode. This list of obsolete data files including structure factors and NMR constraints can be retrieved online from: ***ftp://ftp.rcsb.org/pub/pdb/data/structures/obsolete/***.

h) Sequence – */pub/seqres.txt* – idcode_chainID, molecule type, length of residue sequence, source, sequence. This file can be viewed/retrieved from: ***ftp://ftp.rcsb.org/pub/pdb/derived_data/***.

i) Readme.doc – */pub/readme.doc* – Lists changes recently made or proposed to be made to the PDB CD-ROM set.

j) Changes.doc – */pub/changes.doc* – Lists changes recently made or proposed to be made to the PDB CD-ROM set. Readme.doc and changes.doc are the same file.

k) Document.pdf – */pub/document.pdf* – The PDB CD-ROM documentation is included as an Adobe Acrobat file on disk 1.

4. Software

NOTICE:

The Protein Data Bank does not support any software found on the CD-ROM set. This software has historically been included on the CD-ROM set and is included here for continuity.

Software included on this CD-ROM may require modifications and should be used with caution.

C. Citing the PDB

The contents of PDB are in the public domain, but it is expected that the authors of an entry as well as the PDB be properly cited whenever their work is referred to.

Structures used from the PDB should be cited with the **PDB ID** and the **JRNL** reference.
For example, structure 102L should be referenced as:

PDB ID: 102L
D.W.Heinz,W.A.Baase,F.W.Dahlquist,B.W.Matthews
How Amino-Acid Insertions are Allowed in an Alpha-Helix of T4 Lysozyme.
Nature **361** pp. 561 (1993)

The journal reference for the PDB is:
H.M.Berman, J.Westbrook, Z.Feng, G.Gilliland, T.N.Bhat, H.Weissig, I.N.Shindyalov, P.E.Bourne
The Protein Data Bank
Nucleic Acids Research, **28** pp. 235-242 (2000)

The PDB should also be referenced with the WWW address:
http://www.pdb.org/.

The Brookhaven National Laboratory PDB ceased operation on June 30, 1999.
The original journal reference for the BNL PDB is:
F.C.Bernstein, T.F.Koetzle, G.J.B.Williams, E.F.Meyer Jr, M.D.Brice, J.R.Rodgers, O.Kennard, T.Shimanouchi, M.Tasumi
The Protein Data Bank: a computer-based archival file for macromolecular structures.
J. Mol. Biol. **112** pp. 535-542 (1977)

III. About the PDB

The Protein Data Bank (PDB) is an information portal for researchers and students interested in structural biology. At its center is the PDB archive — the sole international repository for 3-dimensional structure data of biological macromolecules.

The PDB integrates a variety of production-level data and software resources, and shares research results and software. The PDB is dedicated to fostering new scientific advances by providing accurate, consistent, well-annotated 3-D structure data that is delivered in a timely and efficient way to a wide audience.

The contents of the PDB archive contain the structural coordinates and related information about proteins, nucleic acids, and protein-nucleic acid complexes. These structures hold significant promise for the pharmaceutical and biotechnology industries in the search for new drugs and the efforts to understand the mystery of human disease. The understanding of what a structure looks like aids in understanding how it functions.

The PDB is managed by Rutgers, The State University of New Jersey; the San Diego Supercomputer Center at the University of California, San Diego; and the National Institute of Standards and Technology — three members of the Research Collaboratory for Structural Bioinformatics.


THE STATE UNIVERSITY OF NEW JERSEY
RUTGERS


SAN DIEGO SUPERCOMPUTER CENTER


NIST
National Institute of
Standards and Technology