Protein Data Bank

Quarterly Newsletter

Release #74 October 1995

October 1995 CD-ROM Release

3821 full-release atomic coordinate entries

Molecule Type

3469	proteins, peptides, and viruses
80	protein/nucleic acid complexes
260	nucleic acids
12	carbohydrates

Experimental Technique

120	theoretical modeling
490	NMR
3211	diffraction and other

The total size of the atomic coordinate entry database is 1350 Mbytes uncompressed.

Table of Contents

Internet Sites

WWW .... http://www.pdb.bnl.gov
FTP (anonymous) .... ftp.pdb.bnl.gov
Gopher .... gopher.pdb.bnl.gov

What's New at the PDB

In order to make it easier to exchange the structure factors, *i.e.*, the observed experimental data from X-ray crystallographic experiments, the PDB, in close collaboration with a number of macromolecular crystallographers, has developed a proposed standard interchange format for structure factors. This standard is in CIF, which stands for the IUCr-developed "Crystallographic Information File". It was chosen both for simplicity of design and for being clearly self-defining, *i.e.*, that the top of the file contains sufficient information for the rest of the file to be read and understood by either a program or a person. It thus allows for an archival record of the data to be stored for future use. This format will be easy to extend, just by adding additional tokens to the top of the file, as new crystallographic experimental methods are developed.

This project has been lead by Dr. Vivian Stojanoff of the BNL Biology Department, who has worked closely with the Chairperson of IUCr Working Group on Macromolecular CIF, Dr. Paula Fitzgerald, and other members of this committee. Vivian recently presented this standard at the CIF Workshop held during the Annual ACA Meeting in Montreal, Canada (see relevant articles in this Newsletter — one summarizing the Workshop and another presenting details of the format). We encourage all depositors of X-ray structures to submit their structure factors and to do so in this format. In fact, over the years, the PDB has observed that one of the most useful reasons for storing structure factors is for the crystallographer who did the experiment to be able to retrieve his/her own data which may be misplaced in their laboratory. In parallel, it will likely foster a great deal of research into methods of structure determination and validation techniques based on comparison of the final models versus the experimental data. All of the presently deposited structure factors have been converted to this format and are available *via* WWW. In addition, the PDB is providing simple filters which convert between this sfCIF format and other prominent formats, such as those in PROLSQ, TNT, X-PLOR and the CCP4 package.

– Joel L. Sussman

Managing the Archives

This article inaugurates a regular column that will discuss the contents and format of PDB coordinate entries. We plan to use this column to alert users of possible changes in our data representation procedures. We will summarize implementation schedules, major policy decisions, and related items. For example, we regularly receive comments and suggestions for improving both the stored data and access to it. Some of these suggestions have been incorporated into PDB entries, while others are still being evaluated. In the future PDB will also provide data in CIF format using the mmCIF data dictionary. These and other changes in PDB operations require that we keep our users abreast of the issues — this column is intended to do just that.

The PDB is pleased to announce the appointment of Nancy Oeder Manning as Outreach Program Coordinator. Included in her responsibilities is the task of making sure that questions regarding entry content and format are resolved. If you would like to discuss the issues, please send e-mail to Nancy at **oeder@bnl.gov** and/or Enrique Abola at **abola1@bnl.gov**. You may also send a message to a larger audience *via* our listserver **(pdb-l@pdb.pdb.bnl.gov)**.

We suggest keeping a close watch on the PDB Format Description document available through our Internet sites (WWW, FTP, and Gopher). This document will reflect changes in PDB Format and will serve as the reference for all questions regarding the content and format of PDB entries.

A common concern is the time frame for incorporating announced changes into existing entries. There are several reasons for the delay in implementation of changes. The first of these was the need to formulate a clear policy on how changes to the entry content and data format are to be introduced and implemented. We have developed the protocol detailed in the following article. The policy will become effective after a short period of public discussion.

– Proposed Changes

Professor George Sheldrick, University of Göttingen, Germany, suggested that we use B-equivalent in ATOM/HETATM records instead of U-equivalent when anisotropic temperature factors are included in an entry. This will obviate the need to check if ANISOU records are present before interpreting the contents of the B-value field. A number of entries recently deposited have anisotropic thermal parameters. Some of these resulted from studies by Keith Wilson and collaborators on very high resolution structures. George Sheldrick also deposited a structure, refined by SHELX, that includes anisotropic thermal parameters and he requested that the B-value field contain B-equivalent instead of U-equivalent.

We acknowledge the merit in this suggestion and do not foresee any major impediments to its adoption. Since there are only a few entries containing ANISOU values that will be affected by this change, at the request of Professor Sheldrick, his entry will be released with the B-equivalent values.

– HET Groups

The number of heterogen groups stored in the PDB is rapidly growing. Users have been requesting improvements in representation and handling of these groups. There are a number of initiatives underway in response to these requests and suggestions. Most important is the collaboration recently initiated with the Cambridge Crystallographic Data Centre (CCDC). As part of this work CCDC has provided the PDB with PreQuest, the processing program they use in preparing new entries for inclusion into their database. This program will help PDB in representing and managing this growing data set.

This software allows us to build 2-D diagrams of the HET groups from the 3-D coordinates. It automatically generates an atom numbering scheme using a standard algorithm. Matching of the 2-D atom names to those used in the crystal study is also provided. It then becomes less important to follow conventions such as IUPAC when naming atoms. IUPAC conventions for atom nomenclature are difficult to use in computer-based applications. Attempts have been made to follow them over the years; however, as users are aware, violations have occurred because of the difficulties involved in representing heterogens.

The PDB will keep users abreast of major developments in this area. Our WWW server **(http://www.pdb.bnl.gov)** contains a link to our heterogen collection and also to other databases that provide additional information about these groups.

– CIF from PDB Entries

The latest macromolecular CIF (mmCIF) dictionary containing definitions for items normally found in PDB entries has recently been released by the working group commissioned by the IUCr. This dictionary uses the CIF standard and is available through PDB's Internet sites (WWW, FTP, and Gopher) for study and comment. PDB is mounting a major effort to read and write CIF coordinate entry files. Frances Bernstein is currently building a table to relate PDB fields to CIF tokens. Once this table is ready it will be shared with the community along with tools that have been developed for handling PDB entries in CIF.

Comments on this column should be sent to Enrique Abola (**abola1@bnl.gov**) or Nancy Oeder Manning **(oeder@bnl.gov)**.

PDB Format Change Policy

The PDB will use the following protocol in making changes to the way PDB coordinate entries are represented and archived. The purpose of the new policy is to allow ample time for everyone to understand these changes and to assess their impact on existing programs. These modifications are necessary to address the changing needs of our users as well as the changing nature of the data that is archived.

1. Comments and suggestions will be solicited from the community on specific problems and data representation issues as they arise.

2. Proposed format changes will be disseminated through the PDB Listserver **(pdb-l@pdb.pdb.bnl.gov)** and PDB's Internet sites (WWW, FTP, and Gopher). They will also be summarized in the PDB Quarterly Newsletter.

3. A sixty-day discussion period will follow the announcement of proposed changes. Comments and suggestions must be received within this time period. Major changes which are not upwardly compatible will be allotted up to twice the standard amount of discussion time.

4. This sixty-day discussion period will be followed by a thirty-day period in which the PDB staff, the PDB Advisory Board, and the User Group Chair will evaluate and reconcile all suggestions. The final decision pertaining to the format change, which lies with the Advisory Board Chair, will then be officially announced *via* the PDB Listserver and PDB's Internet sites (WWW, FTP, and Gopher).

5. Implementation will follow official announcement of the format change. Major changes will not appear in PDB files earlier than sixty days after the announcement, allowing sufficient time to modify files and programs.

6. Changes will be released no more than twice a year, unless extraordinary circumstances require action. This will be done only in consultation with the Advisory Board and following the usual ninety-day discussion and evaluation period.

The PDB format has been in use since the late 1970's. A number of groups including the mmCIF Committee have been looking at ways to upgrade both the file content and the interchange format used by PDB. This is clearly needed due to changes in the data that PDB archives, the size of the database itself, and finally, to allow us to use more up-to-date methods for representing and storing biological data.

The PDB plans to be prudent and deliberate in making changes to the current PDB files in order to minimize the need to change existing programs. In particular, we will explore ways and means of ensuring that programs which read the current ATOM/HETATM records can continue to do so in the foreseeable future.

Finally, we wish to acknowledge Dr. Gerald Selzer of the National Science Foundation who urged us to formulate this policy.

Any questions, comments, or suggestions should be sent to Joel Sussman **(jls@bnl.gov)**, Enrique Abola **(abola1@bnl.gov)**, or Dave Stampf **(drs@bnl.gov)**.

To be Ala or Not to be Ala? That is the Question.

The SEQRES records in each PDB entry represent the complete sequence of the molecule studied. The sequence that appears on ATOM records must match the SEQRES records. It is essential that each depositor answer the sequence-related questions in the PDB Electronic Deposition Form.

One issue that continues to cause problems is addressed in the Deposition Form but often not handled correctly in the coordinate files that are actually sent to the PDB. Following is text taken from the Deposition Form:

> Some procedures force users to identify residues as ALA when side-chain atoms are missing from the density. However, the correct residue name must be used in your deposition, both in the coordinate and SEQRES records.

It is important to recognize the difference between a residue that is identified as ALA because it actually is alanine and a residue for which the side chain atoms could not be located in the electron density maps. In the latter case the residue must be given the correct residue name. When these instructions are not followed, we have a great deal of difficulty in determining whether a residue that does not match the sequence database is a mutation that we were not told about, if it is a misnamed ALA (for example), or if there is some other reason for the mismatch.

In a similar vein, there is a distinction that must be made between the following possible situations:

1. The atoms of residues 45 - 50 could not be located in the experiment. In this case residues 45 - 50 will be included in the SEQRES records although no coordinates are present in the PDB entry.

2. A recombinant experiment was done so residues 45 - 50 are not present in the protein and residue 51 forms a normal peptide bond with residue 44. In this case the SEQRES records will not include residues 45 - 50 because they were not present.

3. Residues 45 - 50 are excised from the protein and there are now two chains instead of one. In this case the protein will be presented as two separate chains and residues 45 - 50 will not appear in the PDB entry.

Providing the necessary information on the Deposition Form will allow us to prepare entries for your structure more quickly and, most importantly, correctly.

PDB Structure Factor Files in CIF – A Proposal

A standard ASCII interchange format for structure factor files based on the CIF format has been developed by the PDB in collaboration with several macromolecular crystallographers. The work was presented by its primary developer, Dr. Vivian Stojanoff of the BNL Biology Department at the CIF Workshop held during the ACA Annual Meeting in Montreal, Canada. The importance of a standard for the submission and retrieval of structure factor data as well as its role in structure validation procedures was discussed.

Structure factor files are received by the PDB for archiving and redistribution in a variety of formats. These files have to be converted by users to the different data input formats required by analysis programs. The purpose of the new standard is to provide a unique submission and retrieval format for structure factors from the PDB. In addition, it will facilitate the interchange of data between laboratories. The proposed standard is in CIF and was prepared in consultation with Dr. Paula Fitzgerald, Chairperson of the IUCr Working Group on Macromolecular CIF. A number of the data names presented should be considered provisional, as they have not yet been formally approved by the IUCr. However, no substantial changes are anticipated in this part of the mmCIF dictionary.

The proposed standard is compatible with most crystallographic packages, such as CCP4, PROLSQ, TNT, X-PLOR, SHELXL93, SFTools, *etc*. A parser (filter) which converts this format into other standard formats will be distributed by the PDB. The examples presented below illustrate the proposed format. Included in the standard are provisions to accommodate a variety of data items from experiments such as multiple wavelength anomalous scattering. A full description and other examples may be accessed through PDB's Internet sites (WWW, FTP, and Gopher).

Comments and suggestions are welcome; please send them to Enrique Abola **(abola1@bnl.gov)**.

– Examples of SF Files in CIF

• First Example

This example is one of the simpler possible cases, in which the depositor has provided measured and calculated structure factors and calculated phases.

Those reflections that have _refln.status set to 'o' were considered observed and were used in refinement. The reflection with _refln.status set to 'f' indicates that the reflection was considered observed but has been excluded from refinement for the purposes of calculating an R-free value.

```
data_r1aaksf

loop_
_database_2.database_code
_database_2.database_id
pdb_sf 'R1AAKSF'
pdb_coords '1AAK'

_audit.creation_date          '1992-04-08'
loop_
_audit_author.name
'COOK,W.J.'
'JEFFREY,L.C.'
'SULLIVAN,M.L.'
'VIERSTRA,R.D.'

#COMPND    UBIQUITIN CONJUGATING ENZYME (E.C.6.3.2.19)
#SOURCE    (ARABIDOPSIS THALIANA)
#AUTHOR    W.J.COOK,L.C.JEFFREY,M.L.SULLIVAN,R.D.VIERSTRA
#REMARK    2
#REMARK    2    RESOLUTION. 2.4 ANGSTROMS.
#REMARK    3
#REMARK    3    REFINEMENT.
#REMARK    3    PROGRAM 1                X-PLOR
#REMARK    3    AUTHORS 1                BRUNGER
#REMARK    3    PROGRAM 2                PROLSQ
#REMARK    3    AUTHORS 2                KONNERT,HENDRICKSON
#REMARK    3    R VALUE                  0.22
#REMARK    3    RMSD BOND DISTANCES      0.025 ANGSTROMS
#REMARK    3    RMSD BOND ANGLES         2.8 DEGREES
#CRYST1    41.800  44.900  83.200  90.00  90.00  90.00  P 21 21 21 4

loop_
_refln.index_h
_refln.index_k
_refln.index_l
_refln.F_meas_au
_refln.F_calc_au
_refln.phase_calc
_refln.status
0     0    18    553.600    538.455    180.00    o
0     0    20    783.700    781.204    180.00    o
.
.
.
17    3     4     21.100    145.713     23.03    o
17    3     5     63.800     81.747    291.83    f
17    4     0     52.000     40.654    270.02    o
```

• Second Example

In this example, the depositor has provided measured and calculated structure factors, calculated phases, measured sigmas, and three sets of heavy-atom derivative data.

```
data_r1bsrsf

loop_
_database_2.database_code
_database_2.database_id
pdb_sf 'R1BSRSF'
pdb_coords '1BSR'

_audit.creation_date'1993-04-28'
loop_
_audit_author.name
'MAZZARELLA,L.'

#COMPND    RIBONUCLEASE (BOVINE, SEMINAL) (BS-RNASE)
#SOURCE    BOVINE (BOS TAURUS) SEMINAL FLUID
#AUTHOR    L.MAZZARELLA
#REMARK    2
#REMARK    2    RESOLUTION. 1.9 ANGSTROMS.
#REMARK    3
#REMARK    3    REFINEMENT.
#REMARK    3    PROGRAM                  X-PLOR
#REMARK    3    AUTHORS                  BRUNGER
#REMARK    3    R VALUE                  0.177
#REMARK    3    RMSD BOND DISTANCES      0.020 ANGSTROMS
#REMARK    3    RMSD BOND ANGLES         3.70 DEGREES
#CRYST1    36.500  66.700  107.500  90.00  90.00  90.00  P 2  21 21 4

loop_
_refln.index_h
_refln.index_k
_refln.index_l
_refln.F_meas_au
_refln.F_calc_au
_refln.phase_calc
_refln.F_meas_sigma_au
_refln.status
 0   0    4    197.8    1351.0    -180.0     1.9    o
 0   0    6    733.2     807.0       0.0     7.0    o
.
.
.
19   3    1    157.0     131.2    -145.0    24.2    o
19   3    2    207.6     199.7     -11.8    23.5    o
19   3    4    273.8     374.2      64.1    18.5    o
```

```
loop_
_phasing_mir.der_id
_phasing_mir.der_details
HgCl4 'best derivative'
AgNO3_1 'silver nitrate at low concentration - two sites'
AgNO3_2 'silver nitrate at high concentration - six sites'

loop_
_phasing_mir_refln.index_h
_phasing_mir_refln.index_k
_phasing_mir_refln.index_l
_phasing_mir_refln.der_id
_phasing_mir_refln.F_meas_au
_phasing_mir_refln.F_calc_au
_phasing_mir_refln.phase_calc
_phasing_mir_refln.F_meas_sigma_au
0    0    4    HgCl4      197.8    1351.0    -180.0    1.9
0    0    4    AgNO3_1    206.7    1462.0    -180.0   12.3
0    0    4    AgNO3_2    367.3    1551.0    -180.0   36.7
0    0    6    HgCl4      733.2     807.0       0.0    7.0
0    0    6    AgNO3_2    856.6     912.0       0.0   15.4
     .
     .
     .
19   3    1    HgCl4      157.0    131.2    -145.0   24.2
19   3    1    AgNO3_1    142.2    156.3    -112.0   12.7
19   3    1    AgNO3_2    168.3    142.3    -162.0   29.6
19   3    2    AgNO3_1    207.6    199.7     -11.8   23.5
19   3    2    AgNO3_2    183.4    201.9      15.0   25.2
```

DALI – Server for Protein Structure Database Searches in 3D

This article was written by Chris Sander, European Molecular Biology Laboratory – European Bioinformatics Institute, Heidelberg, Germany. It describes a service which may be useful to PDB users.

A database search service allowing exploration of the PDB for structural similarities is now available on WWW. Developed by Liisa Holm and Chris Sander of the European Molecular Biology Laboratory – European Bioinformatics Institute (EMBL-EBI) in Heidelberg, with help and input from Antoine Daruvar and Reinhard Schneider, the server takes protein structure coordinates and returns a list of all protein structures judged as similar by the Dali method. Typically, users may want to:

- establish whether a newly determined structure has a unique fold
- find structural homologues not detectable at the sequence level
- explore the spectrum of currently known protein folds
- extract remotely homologous structures for testing of threading methods

To request a search, go to the PDB WWW site or go directly to **http://www.embl-heidelberg.de/dali/dali.html** and follow instructions. The currently preferred mechanism for submission of search coordinates is *via* e-mail to **dali@embl-heidelberg.de**.

If a user is interested in knowing the structural neighbors of a PDB protein complete with alignment information or would like to see the tree organization of all known structures, the Dali search server, which is linked to the FSSP database [L. Holm and C. Sander, *Nucleic Acids Res.* **22**, 3600-9 (1994); L. Holm and C. Sander, *J. Mol. Biol.* **233**, 123-38 (1993); L. Holm and C. Sander, *TIBS* in press (Sept. 1995)], holds the results of a continuously updated all-against-all comparison of known protein structures in the PDB. For each protein in the PDB, its structural neighbors can be retrieved, complete with alignment information, and, if needed, protein coordinates optimally superimposed in a common frame of reference.

If a user would like to inspect sequence families grouped around each protein structure, the Dali server makes use of the continuously updated HSSP database of PDB/Swissprot alignments [C. Sander and R. Schneider, *Nucleic Acids Res.* **22**, 3597-9,1994)]. For example, the Swissprot sequence families (aligned using sequence alignment) of two remote homologues (aligned in 3D) can be retrieved phased on the 3D alignment, *e.g.*, the sequence families of actin and hsp70.

For further information, contact Liisa Holm at the EMBL-EBI **(holm@embl-ebi.ac.uk)**.

WHATCHECK – Tool for Verification of Protein Structures

This article was written by Rob Hooft, European Molecular Biology Laboratory, Heidelberg, Germany. It describes a tool which may be useful to PDB users.

A new stand-alone version of the protein structure checking tools developed as part of the WHAT IF modeling system is now available for downloading. The stand-alone version, called WHATCHECK, was developed by Rob Hooft and Gerrit Vriend of the European Molecular Biology Laboratory (EMBL) in Heidelberg and is being made freely available to academic and commercial users alike. Crystallographers, NMR spectroscopists, and modelers interested in using the software in their own lab should free 80 MBytes of disk space on a Silicon Graphics machine and download executable and source code *via* anonymous FTP from Heidelberg (**swift.embl-heidelberg.de/whatcheck**), from the PDB *via* anonymous FTP (**ftp.pdb.bnl.gov/pub/whatcheck**), or from the PDB WWW site **(http://www.pdb.bnl.gov)**. Versions for other systems may be made available if there is sufficient demand.

The methods in WHATCHECK are based on several years of experience analyzing protein models and include the following:

- analysis of packing quality
- analysis of backbone conformation and rotamers using a position-specific residue database

- analysis of the hydrogen bond network, resulting in corrections to HIS/GLN/ASN side-chain orientations and HISD/HISE/HISH assignments
- temperature factor analysis
- water molecule position checks
- IUPAC and IUCr convention checks
- a number of geometric checks

WHATCHECK functionality is also part of the Biotech Validation server available from EBI, PDB, and EMBL.

Questions or suggestions regarding WHATCHECK should be directed *via* e-mail to **Rob.Hooft@EMBL-Heidelberg.DE**.

Biotech Protein Structure Validation Server Now On-Line

This article was written by Chris Sander, European Molecular Biology Laboratory – European Bioinformatics Institute, Heidelberg, Germany. It describes a service which may be useful to PDB users.

The Commission of the European Union-funded Biotech partners are:

Shoshana Wodak, Joan Pontius, Alexei Vagin, Jean Richelle
Universitè Libre, Bruxelles, BE

Keith Wilson, Victor Lamzin
EMBL-HH, Hamburg, DE

Chris Sander, Rob Hooft, Gerrit Vriend, Micheal Scharf
EMBL-HD, Heidelberg, DE

Janet Thornton, Roman Laskowski, Malcolm MacArthur
University College London, UK

Rob Kaptein, Ton Rullmann, Jurgen Doreleijers
Bijvoet Center, Universiteit Utrecht, NL

Eleanor Dodson, Gideon Davies, Jan Zelinka, Garib Murshudov
University of York, UK

These partners, in collaboration with the EBI Outstation of the European Molecular Biology Laboratory (EMBL-EBI) and the Protein Data Bank, Brookhaven National Laboratory (PDB, BNL), announce a new WWW service for validating protein structures. The server takes a set of protein coordinates in PDB format and returns a set of check reports that evaluate various stereochemical, geometric, and physical properties. The evaluation is based on careful and systematic comparison with values derived from a database of well-determined structures. These checks may be particularly useful to crystallographers and NMR spectroscopists in the final stages of structure determination prior to submission of model coordinates to the public databases. They may also be used to check protein models built by theoretical methods, such as homology modeling.

The service is now available from three sites:

European Bioinformatics Institute
EMBL-EBI, Hinxton Hall, Cambridge, UK
http://saturn.embl-ebi.ac.uk:8400/
or **http://www.embl-ebi.ac.uk/**

Protein Data Bank
Brookhaven National Laboratory
Upton, New York, USA
http://biotech.pdb.bnl.gov:8400/
or **http://www.pdb.bnl.gov/**

European Molecular Biology Laboratory
EMBL-HD, Heidelberg, DE
http://www.embl-heidelberg.de:8400/
or **http://www.sander.embl-heidelberg.de/**

For further information, contact Chris Sander at the EMBL-EBI (**Chris.Sander@embl-heidelberg.de**).

Notes of a Protein Crystallographer

– Professor M. G. Replacement's 65th Birthday

*This article was written by Cele Abad-Zapatero, Abbott Laboratories, Abbott Park, IL, USA. He intends to contribute regularly under this heading. If you have comments or suggestions please contact him at **abad@abbott.com**.*

Nowadays, it is possible that many older practitioners in the field of macromolecular crystallography, or even some novices, experience a sense of *deja vu* when reading the methodology section of many crystallographic papers. They have encountered many times sentences like: “the structure was solved by the method of Molecular Replacement as implemented in the program suite [...]....”. Or, in a different context: “the initial phases were improved by non-crystallographic electron density averaging between the N copies in the asymmetric unit [...]”. Yet the ideas, concepts, and even the terminology were *terra incognita* only thirty years ago. I do not intend to explain the method here, nor do I intend to statistically show how many macromolecular structures have been solved by the explicit or implicit use of those ideas. I just want to pay homage to the person who, approximately thirty years ago, had the intellectual vision of using the conservation of protein folds and the ‘redundancy’ of information in the asymmetric unit to aid in the structure solution of macromolecular structures.

Professor Molecular G. Replacement was born in Frankfurt, Germany on July 30, 1930 and emigrated to England with a member of his family when he was nine years old. He obtained his B.Sc. from the University of London in 1950, M.Sc. in 1953 and later a Ph.D. in Chemistry from the University of Glasgow. He did his postdoctoral research from 1956 through 1958 with Professor W. Libscomb in Minneapolis. Inspired by a lecture by Dorothy Hodgkin to work on the determination of the crystal structure of biological macromolecules, he returned to England to work with Max Perutz on the structure of haemoglobin at the MRC during the exciting years spanning from 1958 to 1964. From then on, his ideas, programs, structures, and papers have had a tremendous influence in the field of macromolecular crystallography.

He married a remarkable woman, Audrey Pearson, in 1954 and they have three children (Alice, Martin, and Heather). Many old-timer crystallographers will remember the very first "cartoon" sketches of lactate dehydrogenase (LDH). Following Anders Liljas' suggestion, they were drawn by Audrey to simplify the wanderings of the polypeptide chain in space. Those drawings that today are so commonplace in all kinds of colors, shades and hues, have their origin in those crude hand-drawn diagrams. Incidently, she is also a superb potter.

Some may wonder, why this sudden interest in recognizing Professor M. G. Replacement. I must confess that his 65th birthday is only partly the reason. The idea originated when I recently saw a paper in one of the leading journals of our field where a structure had been solved by phase refinement using electron density averaging and there was no reference to Professor M. G. Replacement. At first, thinking it was pretty sad, I decided to write this note. But then, on second thought, I realized that in the end it may be of great honor to be passed on into oblivion. Do we quote Sir Isaac Newton every time we use the principle of inertia? Do we refer to the author of the *Principia* every time we use any of his equations? Certainly not. His work is all ingrained within the fabric of our science and our culture. Similarly, the work of Professor M. G. Replacement is now, and for evermore, part of the framework of macromolecular crystallography.

Taking some literary liberties, I would like to finish this brief homage with an adaptation of a well known American folk song. I leave it to the reader to find out the first and last names of Professor M. G. Replacement, his favorite hobby (I already disclosed the name of his wife and partner), and the name of the river that crosses the city in Indiana were both have lived since 1964.

M... sail the boat ashore, Hallelujah
M... sail the boat ashore, Hallelujah

Audrey help to trim the sails, Hallelujah
Audrey help to trim the sails, Hallelujah

Wabash river is chilly and cold, Hallelujah
Freezes the body but not the soul, Hallelujah

Wabash river is deep and wide, Hallelujah
Milk and honey on the other side, Hallelujah

CIF Workshop at 1995 Annual ACA Meeting

This article was written by Philip Bourne, San Diego Supercomputer Center (SDSC), San Diego, CA, USA.

The Crystallographic Information File (CIF) is a data representation and exchange format commonly used in small molecule crystallography. Most of the small molecule solution packages can produce CIF files, a common form of submission for papers to Acta C.

With the availability of a macromolecular CIF (mmCIF) dictionary, a powder diffraction dictionary, and other dictionaries currently in progress (modulated structures, meta graphics, amino acid and nucleotide properties, *etc.*), the need for a Workshop describing these dictionaries and new software which uses them was considered timely. The Workshop was sponsored by COMCIFs, the IUCr-appointed committee which oversees the CIF standard, organized by Phil Bourne from SDSC, and held during the 1995 Annual ACA Meeting held in Montreal, Canada. Approximately forty people were in attendance.

Bourne opened the Workshop with a brief discussion on the history of CIF and described how similar efforts are only now starting to emerge in other disciplines. He emphasized the need for the crystallographic community to recognize that the high level of formalized data description achieved is presently unique among scientific disciplines. This offers new opportunities for software development as well as better access to a fast growing body of data through the use of comprehensive databases.

Syd Hall (University of Western Australia), the founder of CIF, elaborated on the history of the Self-defining Text Archival and Retrieval (STAR) format from which CIF is derived as well as CIF itself. STAR is a simple set of rules from which a Dictionary Definition Language (DDL) can be derived. The DDL defines the form of the various dictionaries and, as Hall pointed out, is critical to the development of good software. He also described CIFtbx (CIF Toolbox), a set of Fortran routines for the basic reading and writing of CIF files (**ftp site 130.95.232.12)**. He emphasized the need for further software development particularly in the reading and browsing of CIF files.

Brian McMahon (IUCr) provided insight into CIF processing at the IUCr offices in Chester, and the software developed to process CIF-based submissions to Acta C. Of the 582 submissions to Acta C this year, 76 percent were CIF's that included the text of the paper. Any submissions in hardcopy form are first turned into a CIF and all submissions undergo data checking and subsequent format conversion into an easily-read form. Once accepted, a paper may automatically be converted into a final typeset version in the style of Acta C. McMahon also reported on three recent developments of great benefit to authors. First, is the availability of a booklet entitled "A Guide to CIF for Authors." Second, is the ability to submit a CIF file *via* e-mail to **checkcif@iucr.ac.uk**. In return, the submitter will receive a detailed report of errors and potential errors found. This includes syntax errors, data items not conforming to the official dictionaries (a potential error), missing data items, data items with unusual values, derived values, and an indication that higher symmetry than reported may exist. Once the entry has passed checkcif, the CIF can be sent to **printcif@iucr.ac.uk**, the third recent development. Printcif produces a PostScript-formatted version of the file with any problematic values highlighted. This formatted version is not used by the journal, but is easy to read and correct.

Paula Fitzgerald (Merck Sharp and Dohme Research), Chairperson of the mmCIF Working Group, presented an overview of the mmCIF dictionary and an update of recent progress. The current mmCIF dictionary is composed of over 20,000 lines and contains a description for several thousand data items organized into over 300 categories. The categories are further subdivided into groups to provide a hierarchical representation which traces the progress of the crystallographic experiment and the subsequent structure. While the development of the dictionary took five years of volunteer effort, it should be acknowledged that the dictionary is comprehensive and provides very useful reference work. The draft mmCIF dictionary is finished except for the incorporation of changes recognized by the community and final editorial checking by COMCIFs. The current draft dictionary can be found as an ASCII file at the WWW site **http://ndbserver.rutgers.edu/mmCIF** or on PDB's WWW home page. Also present are limited examples and introductory material. The contents of the site will be expanded in the near future. A listserver has also been established. To subscribe, send an e-mail message to **mmciflist@ndbserver.rutgers.edu** containing the words: subscribe mmciflist your-name. Version 1.0 of the dictionary will be available within three to six months.

Eldon Ulrich (University of Wisconsin) described progress with the NMR dictionary (NMRif). This project began at the April 10, 1994 Workshop, Biological Macromolecular NMR Data Exchange and Archiving, organized by members of BioMagResBank and the mmCIF committee. The CIF and mmCIF concept and its development was described to scientists from the NMR community and proposed as a format for macromolecular NMR data exchange. BioMagResBank, under the direction of Eldon Ulrich, agreed to undertake the task of developing the dictionary with the help and advice of volunteers from the NMR community. The dictionary is represented as a relational database using a schema design tool called Opossum. Development of the relational schema has been carried out in collaboration with Miron Livny and Yannis Ioannidis, computer scientists at the University of Wisconsin. The current dictionary contains over 260 tables or categories comprised of more than 750 unique data names. While no estimate was given for the dictionary completion date, it is expected to grow to double or triple its current size. It was suggested that Opossum be used as a graphical tool for representing and browsing the formidable mmCIF dictionary.

Gotzon Madariaga (University del Pais Vasco, Bilbao, Spain) presented the main features of the modulated structures dictionary he is developing, which is a superset of all the data items defined by the Commission on Aperiodic Crystals. A draft of this dictionary has been submitted to the IUCr. Interestingly, he raised several issues relating to problems with CIF which had also been noted by the mmCIF developers. Notable was the need to provide a reference between related blocks of data possibly contained in different files. The lack of nested loops in CIF (which are available in STAR) was also raised because the lack of these loops makes it more cumbersome to represent the data.

The final morning session was given by Bourne and concerned two new dictionaries based on CIF. The first was a dictionary to process abstracts for the ACA Meeting. Each of seventy CIF abstracts was submitted either using a WWW form or by completing a template form. The WWW form processing was automatic whereas many of the templates had to be hand-edited. Bourne concluded that since it was not necessary to collect statistics on the CIF submissions, processing them was more effort than was warranted. The second dictionary is a meta graphics (mg) dictionary to facilitate data exchange between graphics programs. The mg dictionary, while preliminary, is designed to capture all the salient features seen in the display of a macromolecule such that the biological structure/function visualized in the graphics representation are preserved and can be loaded into a variety of graphics programs. A long-term advantage of this approach is that graphics images are easily stored in a database and can be extracted as needed.

Vivian Stojanoff (Brookhaven National Laboratory) began the afternoon session with a discussion of an extension to the mmCIF dictionary for representing structure factors. While the current structure dictionary is not considered complete, it does describe more than the basic h,k,l,F and sigmaF found in many existing PDB submissions. Currently there is support for data from multiple derivatives and wavelengths.

John Westbrook (Rutgers University) described the Dictionary Definition Language (DDL) he developed in response to the needs of the macromolecular dictionary. The current small molecule core dictionary uses DDL version 1.4. Westbrook developed version 2.1 which is upwardly compatible with version 1.4 but is much more rigorous. While this has little significance to the crystallographer, it is important to the software developer who needs a consistent way of representing both dictionaries and data files so that one piece of software can be used on both. Version 2.1 of the DDL offers that opportunity.

Syd Hall returned to describe the program Xtal_GX which provides a general approach to processing CIFs and uses the CIFtbx he described earlier in the day. It also has a graphics feature for display purposes.

The final formal presentation was given by Weider Chang (Columbia University) who described the next-generation, object-oriented tools he is developing with Bourne for basic CIF manipulation. Chang presented the basic layout of the library which is written in Objective C and from which browsers, a CIF2HTML convertor, and several dictionary checking tools have been developed. The CIF2HTML convertor was used in the ACA abstract procedure.

A lively discussion followed the formal presentations pertaining to the need of recognizing which subset of mmCIF data items will constitute a formal PDB entry. Attendees seemed well satisfied with the workshop — perhaps it will lead to development of other dictionaries as well as new, useful software.

BROOKHAVEN ORDER FORM

Name of User	______________________	Date	______________
Organization	______________________	Phone	______________
Address	______________________	Fax	______________
	______________________	E-mail	______________
	______________________		

- ***Price is valid through September 30, 1996***
- ***Price is per CD-ROM set released --- releases occur four times per year***
- ***Facsimile and phone orders are not acceptable***

The Protein Data Bank ***MUST*** receive all three of the following items before shipment can be completed (please send all required items together *via* postal mail — facsimile and phone orders are ***NOT*** acceptable):

1. Completed order form;
2. Mailing label indicating exact shipping address;
3. Payment (using one of the two options below):

--> Check payable to Brookhaven National Laboratory in US dollars and drawn on a US bank. Foreign checks cannot be accepted and will be returned.

--> Original purchase order payable to Brookhaven National Laboratory. After your order is processed, you will be invoiced by Brookhaven National Laboratory. Please indicate exact address invoice should be sent to:

____

____

____

____

A wire transfer is acceptable only ***AFTER*** we have received an original purchase order from your organization and you have been invoiced by Brookhaven. After receiving Brookhaven's invoice, your bank may send a wire transfer to:

Bank name	:	Morgan Guaranty Trust Company of New York
Account name	:	Brookhaven National Laboratory
Account number	:	076-51-912

Please send all required items together *via* postal mail to:

Protein Data Bank Orders
Chemistry Department, Building 555
Brookhaven National Laboratory
P.O. Box 5000
Upton, NY 11973-5000
USA

1 Protein Data Bank CD-ROM Set — ISO 9660 Format $332.26

(tax and shipping charges not applicable)

This page intentionally left blank.

AFFILIATED CENTERS

Twenty-two affiliated centers offer DATAPRTP information for distribution. These centers are members of the Protein Data Bank Service Association (PDBSA). Centers designated with an asterisk(*) may distribute DATAPRTP information both on-line and on magnetic or optical media; those without an asterisk are on-line distributors only.

BMERC
BioMolecular Engineering Research Center
College of Engineering, Boston University
Boston, Massachusetts
Nancy Sands (617-353-7123)
sands@darwin.bu.edu
http://bmerc-www.bu.edu/

*BIOSYM
BIOSYM Technologies, Inc.
San Diego, California
Rick Lee (619-546-5536)
rickl@biosym.com
http://www.biosym.com/

BIRKBECK
Crystallography Department
Birkbeck College, University of London
London, United Kingdom
Alan Mills (44-171-6316810)
a.mills@cryst.bbk.ac.uk
http://www.cryst.bbk.ac.uk/PDB/pdb.html/

CAN/SND
Canadian Scientific Numeric Data Base Service
Ottawa, Ontario, Canada
Roger Gough (613-993-3294)
cansnd@vm.nrc.ca

CAOS/CAMM
Dutch National Facility for Computer Assisted Chemistry
Nijmegen, The Netherlands
Jan Noordik (31-80-653386)
noordik@caos.caos.kun.nl
http://www.caos.kun.nl/

*CCDC
Cambridge Crystallographic Data Centre
Cambridge, United Kingdom
David Watson (44-1223-336394)
watson@chemcrys.cam.ac.uk

CSC
CSC Scientific Computing Ltd.
Espoo, Finland
Heikki Lehvaslaiho (358-0-457-2076)
heikki.lehvaslaiho@csc.fi
http://www.csc.fi/

CINECA
NE Italy Interuniversity Computing Center
Casalecchio di Reno (BO), Italy
Laura Setti (39-51-6599478)
asltc0@icineca.cineca.it

ICGEB
International Centre for Genetic Engineering and Biotechnology
Trieste, Italy
Sandor Pongor (39-40-3757300)
pongor@icgeb.trieste.it

EMBL
European Molecular Biology Laboratory
Heidelberg, Germany
Hans Doebbeling (49-6221-387-247)
hans.doebbeling@embl-heidelberg.de
http://www.EMBL-Heidelberg.DE/

INN
Israeli National Node
Weizmann Institute of Science
Rehovot, Israel
Leon Esterman (972-8-343934)
lsestern@weizmann.weizmann.ac.il

*JAICI
Japan Association for International Chemical Information
Tokyo, Japan
Hideaki Chihara (81-3-5978-3608)

*MAG
Molecular Applications Group
Palo Alto, California
Hilary Jensen (415-473-3039)
hilary@suerte.mag.com
http://hyper.stanford.edu/~Mag/

*MSI
Molecular Simulations Inc.
Burlington, Massachusetts
Lance J. Ransom Wright (617-229-9800)
lance@msi.com
http://www.msi.com/

NCHC
National Center for High-Performance Computing
Hsinchu, Taiwan, ROC
Jyh-Shyong Ho (886-35-776085; ex: 342)
c00jsh00@nchc.gov.tw

NCSA
National Center for Supercomputing Applications
University of Illinois at Urbana-Champaign
Champaign, Illinois
Patricia Carlson (217-244-0768)
pcarlson@ncsa.uiuc.edu

NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION
National Library of Medicine
National Institutes of Health
Bethesda, Maryland
Stephen Bryant (301-496-2475)
bryant@ncbi.nlm.nih.gov
http://www.ncbi.nlm.nih.gov/

*OML
Oxford Molecular Ltd.
Oxford, United Kingdom
Steve Gardner (44-1865-784600)
sgardner@oxmol.co.uk
http://www.oxmol.co.uk/

*OSAKA UNIVERSITY
Institute for Protein Research
Osaka, Japan
Yoshiki Matsuura (81-6-879-8605)
matsuura@protein.osaka-u.ac.jp

PITTSBURGH SUPERCOMPUTING CENTER
Pittsburgh, Pennsylvania
Hugh Nicholas (412-268-4960)
nicholas@psc.edu
http://pscinfo.psc.edu/biomed/biomed.html/

SEQNET
Daresbury Laboratory
Warrington, United Kingdom
User Interface Group (44-1925-603351)
uig@daresbury.ac.uk

*TRIPOS
Tripos, Inc.
St. Louis, Missouri
Akbar Nayeem (314-647-1099; ex: 3224)
akbar@tripos.com

Protein Data Bank
Chemistry Department, Bldg. 555
Brookhaven National Laboratory
P.O. Box 5000
Upton, NY 11973-5000
USA

Statement of Support

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

Contacts

Telephone 516-282-3629
Facsimile 516-282-5751

Internet:
- pdb@bnl.gov .... general correspondence
- orders@pdb.pdb.bnl.gov .... order information
- sysadmin@pdb.pdb.bnl.gov .... network services
- listserv@pdb.pdb.bnl.gov .... Listserver subscriptions
- pdb-l@pdb.pdb.bnl.gov .... Listserver postings
- errata@pdb.pdb.bnl.gov .... entry error reporting

Please include your name, postal mailing address, e-mail address, facsimile

Staff

Joel L. Sussman, Head
David R. Stampf, Sr. Project Mgr.
Enrique E. Abola, Science Coordinator
Jaime Prilusky, Interim Head Database Dev.

Frances C. Bernstein
Judith A. Callaway
Minette Cummings
Betty R. Deroski
Pamela A. Esposito
Arthur Forman
Patricia A. Langdon
Michael D. Libeson
Nancy O. Manning
John E. McCarthy
Janet L. Sikora
Regina K. Shea
Karen E. Smith
Dejun Xue

Internet Sites

WWW .... http://www.pdb.bnl.gov
FTP (anonymous) .... ftp.pdb.bnl.gov
Gopher .... gopher.pdb.bnl.gov

First Class