Number 1

September '74

Protein Data Bank

Newsletter

We thought it would be a good idea to mail out a report on the status of the Protein Data Bank and to establish at this time a regular newsletter. Some of the material in this first edition of the newsletter may be familiar to you. We promise that the next edition will be much smaller!

Deposition of Coordinates

Data may be deposited by filling out the form in Appendix 1. Tape or cards rather than a listing are appreciated. Mail these to

T.F. Koetzle
Department of Chemistry
Brookhaven National Laboratory
Upton, New York 11973
Telephone: 516-345-4384

Coordinate Directory

The coordinate sets in final distributable form are listed below along with coordinate sets soon to be available (marked *):

carboxypeptidase A
carp muscle calcium binding parvalbumin
α-chymotrypsin
cytochrome b_5
flavodoxin *
D-glyceraldehyde-3-phosphate dehydrogenase *
horse hemoglobin (deoxy and met)
lactate dehydrogenase
lamprey hemoglobin
lysozyme *
myoglobin
pancreatic trypsin inhibitor
papain
rubredoxin
staphylococcal nuclease
subtilisin
thermolysin*

Format

The format of the coordinates is given in Appendix 2. Torsion angles, structure factors and phases are also available for some proteins, as indicated in Appendix 3.

Access and Distribution

1. Send in request form in Appendix 4.
2. Microfiche describing each protein are currently being prepared by Richard Feldmann of NIH. The details of distribution are being ironed out now.
3. Computer access
   a) It is possible for users without accounts at BNL to print the files if they have a teletype with an acoustic coupler.

      Dial 516-345-2210 (300 baud)
      516-345-4191 (100 baud)

      Log in by typing ON PDB C999 C/R. (No jobs may be submitted to the Brookhaven CDC 6600 on this particular account). List the names of available files by typing LIST C/R, and answering the queries which follow by typing NO C/R. List the contents of any file by typing OUTPUT NAME C/R. There may be a delay of a few minutes while the file is loaded from tape. When finished, log off by typing OFF C/R, and answering the query, "EMPTY OR T OR INHIBITED C FILES WILL BE LOST" by typing YES C/R.

   b) If you wish to establish an account at BNL contact T. Koetzle. It will then be possible to search for fragments of structures using the program SEARCH (Appendix 5) or with any other program of the user's choosing.

Mini-files

Mini files are being assembled under account PDB which contain coordinates of particular parts of proteins. For example, files have been created with the coordinates of the atoms within 8 Å of the Fe atom in heme proteins.

INPUT TO PROTEIN DATA BANK

1. Name of Protein

2. Name and Address of Contributor

3. Names of Co-Authors

4. Submitted Data Atomic Coordinates ☐ Structure Factors ☐
   Torsion Angles ☐ Electron Densities ☐

5. Magnetic Tape Specifications

   7-track ☐ 556 bpi ☐ BCDIC ☐ Unlabelled ☐

   9-track ☐ 800 bpi ☐ EBCDIC ☐ Labelled ☐

   1600 bpi ☐

   Other Details:

6. Unit Cell Data a = α = Space Group

   b = β = Z =

   c = γ =

7. Footnotes for Specific Atoms or Residues

8. Format of Atomic Coordinates

Real Space Refinement Program (Diamond) ☐

Other:

9. Transformation of Atomic Coordinates

If the submitted coordinates are not expressed as fractions of the unit cell edges then state the transformation necessary to obtain fractional cell coordinates (xyz).

10. Format of Torsion Angles

Real Space Refinement Program (Diamond) ☐

Other:

11. Format of Structure Factors

12. Format of Electron Densities

13. General Remarks (inc. Literature References)

ATOMIC COORDINATE AND TORSION ANGLE FILES

For each protein data set the file consisists of records each of 132 characters.

The record sequence is as follows:-

COMPND : Name of protein
AUTHOR : Names of contributor and co-authors
CRYST1 : Unit cell data
DECODE : List of element types present in protein
REMARK : General remarks
FTNOTE : Footnotes relating to specific atoms or residues
ORIGX1-3: Transformation matrix (fractional cell coords. → submitted coords.)
SCALE1-3: Transformation matrix (fractional cell coords. → orthogonal Å coords.)
Atomic Coordinate Records
Torsion Angle Records
End of Data Record

The first eight record types (COMPND to SCALE) are indicated by the first six characters of the record.

Atomic coordinate and torsion angle records are patterned after the Real Space Refinement Program of R. Diamond.

Each protein is assigned an identification code and this code is carried on all records for the data set, except atomic coordinate, torsion angle and end of data records. The code consists of six letters and a possible two numeric digits. The latter are provided to distinguish multiple data sets for the same protein.

In describing record formats it will be convenient to use the punched-card analogy and refer to column numbers.

LIST OF PROTEIN DATA BANK HOLDINGS

AC: Atomic Coordinates SF: Structure Factors AD: Available for Distribution

TA: Torsion Angles ED: Electron Densities

Name of Protein	Contributor	AC	TA	SF	ED	AD
carboxypeptidase A	Lipscomb	X				X
α-chymotrypsin	Blow	X		X		X
cytochrome b_5	Mathews	X				X
lactate dehydrogenase	Rossmann	X	X			X
lamprey hemoglobin	Hendrickson and Love	X	X	X		X
pancreatic trypsin inhibitor	Huber	X				X
subtilisin	Kraut	X				X
myoglobin	Watson	X				X
rubredoxin	Jensen	X				X
torsion angles for 11 proteins (see J. Mol. Biol. 75, 13 (1973))	Wu and Kabat		X			X
horse deoxyhemoglobin	Perutz	X				X
horse methemoglobin	Perutz	X				X
papain	Drenth	X	X			X

7. ORIGX1-3

Format	6A1	F10.5	F10.5	F10.5	F10.5	8A1
Cols.	1-6	11-20	21-30	31-40	41-50	73-80
	ORIGX1	O_{11}	O_{12}	O_{13}	T_1	Id. Code
	ORIGX2	O_{21}	O_{22}	O_{23}	T_2	Id. Code
	ORIGX3	O_{31}	O_{32}	O_{33}	T_3	Id. Code

Note:- Let the original submitted coordinates be $X_oY_oZ_o$
Let the fractional cell coordinates be x y z
Then $X_o = O_{11}x + O_{12}y + O_{13}z + T_1$
$Y_o = O_{21}x + O_{22}y + O_{23}z + T_2$
$Z_o = O_{31}x + O_{32}y + O_{33}z + T_3$

8. SCALE1-3

Format	6A1	F10.5	F10.5	F10.5	8A1
Cols.	1-6	11-20	21-30	31-40	73-80
	SCALE1	S_{11}	S_{12}	S_{13}	Id. Code
	SCALE2	S_{21}	S_{22}	S_{23}	Id. Code
	SCALE3	S_{31}	S_{32}	S_{33}	Id. Code

Note:- Let the orthogonal Å coordinates be X Y Z
Let the fractional cell coordinates be x y z
Then $X = S_{11}x + S_{12}y + S_{13}z$
$Y = S_{21}x + S_{22}y + S_{23}z$
$Z = S_{31}x + S_{32}y + S_{33}z$

The orthogonal cell ($\underset{\sim}{A},\underset{\sim}{B},\underset{\sim}{C}$) is related to the crystal cell ($\underset{\sim}{a},\underset{\sim}{b},\underset{\sim}{c},$) as follows:-
$\underset{\sim}{A}$ is parallel to $\underset{\sim}{a}$; $\underset{\sim}{B}$ is parallel to $\underset{\sim}{C} \times \underset{\sim}{A}$; $\underset{\sim}{C}$ is parallel to $\underset{\sim}{a} \times \underset{\sim}{b}$.

9. Atomic Coordinate Record

Field	Cols.		
1	1-10	Fractional x-Coordinate	(F10.5)
2	11-20	Fractional y-Coordinate	(F10.5)
3	21-30	Fractional z-Coordinate	(F10.5)
4	31-40	Atomic Radius	(F10.2)
5	41-45	Atom Type Number	(I5)
6	46-50	Sequence Number	(I5)
7	51-55	Atomic Coordinate Record Number	(I5)
8	56-64	Electron Count	(F9.4)
9	65-68	Residue Name	(1X,A3)
10	69-80	Residue Identifier; Atom Identifier	(A6,A6)
11	81-90	Orthogonal x-Coordinate	(F10.5)
12	91-100	Orthogonal y-Coordinate	(F10.5)
13	101-110	Orthogonal z-Coordinate	(F10.5)
14	111-120	Retrieval Code Number	(10A1)
15	121-124	Occupancy Factor	(1X,F3.1)
16	125-129	Temperature Factor	(1X,F4.1)
17	130-132	Footnote Number	(1X,I2)

Notes:-

Field 4: This field may contain a value for the effective atomic radius; otherwise it is blank or zero.

Field 5: The atom type number is an integer > 0.
These numbers correspond to the order in which the element symbols appear in the DECODE record.

Field 6: The sequence number is normally an integer > 0.
A value of 0 indicates a dummy atom or a chain terminator.

Field 7: The atomic coordinate record number is simply a serialisation number. Thus it is 1 for the 1st record, 2 for the 2nd etc. etc.

Field 8: This field may contain a value for the electron count; otherwise it is blank or zero.

Field 9: The residue name is indicated by a standard abbreviation e.g. GLY. The first character of the field must be blank. For a list of residue names see Appendix 1.
Diamond's program requires that the amide N atom be named as part of a peptide unit. Hence the name of the amino-acid residue will be placed on the first atom following the amide N (if it occurs first, as is usually the case). Note that the sequence number (field 6) refers to the amino acid residue and not the peptide unit. When the structure consists of several independent chains then each chain terminates with TER in field 9; record numbers will follow sequentially.

Field 10: The residue identifier (A6) normally has a leading blank character. It is usually the same as the sequence number (field 6).
The atom identifier (A6) normally has a leading blank character. The conventions for atom identifiers are given in Appendix 2.

Fields 11-13: These fields carry the orthogonal Å coords. generated from the fractional cell coords. (fields 1-3) by the application of the matrix SCALE.

Field 14: The retrieval code number is described in Appendix 2.

Field 15: If the occupancy factor field is blank a value of 1.0 is assumed.

Field 16: If the temperature factor field is blank a value of 0.0 is assumed.

Field 17: A number in this field indicates the presence of an associated footnote.

10. Torsion Angle Record

Field	Cols.		
1-3	1-30	Blank	(30X)
4	31-40	Torsion Angle Value (in degrees)	(F10.5)
5	41-45	"Atom Type" Number	(I5)
6	46-50	"Sequence" Number	(I5)
7	51-55	Torsion Angle Record Number	(I5)
8	56-64	Elastic Constant	(F9.4)
9	65-68	Blank	(4X)
10	69-80	Residue Identifier; Torsion Angle Identifier	(A6,A6)
11-16	81-132	Blank	(52X)

Notes:-

- Torsion angle records have the same overall format as atomic co-ordinate records and are easily distinguished by the sign of the number in field 5 - negative for torsion angles and positive for atomic coordinates.
  Torsion angle records are usually interleaved among the atomic coord. records.
  Field 5: The "atom type" number is an integer > 0.
  Field 6: For a main-chain torsion angle this is an integer > 0.
  For a side-chain torsion angle this is an integer < 0.
  Field 7: This is the serialisation number, as for atomic coordinate records.
  Field 8: This field may contain an elastic constant, otherwise it is blank or zero.
  Field 10: The residue identifier (A6) normally has a leading blank character. The torsion angle identifier (A6) does not have a leading blank character, i.e. the angle name (e.g. CHI,PHI) begins in col. 75.

11. End of Data Record

Field	Cols.		
1-4	1-40	Blank	(40X)
5	41-45	1	(I5)
6	46-50	-1	(I5)
7-8	51-64	Blank	(14X)
9	65-68	END	(1X,A3)
10-16	69-132	Blank	(64X)

Note:- The final record of the data set takes the general format of an atomic coordinate record.

APPENDIX 1

Residue Names, Abbreviations, Types, Identification Numbers

Residue	Abb.	Type	No.	Residue	Abb.	Type	No.
Alanine	ALA	1	2	Isoleucine	ILE	1	5
β-Alanine	ALB	1	25	Leucine	LEU	1	4
γ-Aminobutyric acid	ABU	1	26	Lysine	LYS	4	12
Arginine	ARG	4	15	Methionine	MET	1	13
Asparagine	ASN	5	9	Ornithine	ORN	4	30
Aspartic acid	ASP	3	8	Phenylalanine	PHE	1	16
Betaine	BET	4	28	Proline	PRO	1	19
Cysteine	CYS	6	20	Pyrollidone carboxylic acid	PCA	5	32
Cystine	CYS	6	21	Sarcosine	SAR	1	27
Glutamic acid	GLU	3	10	Serine	SER	2	6
Glutamine	GLN	5	11	Taurine	TAU	3	31
Glycine	GLY	7	1	Terminator	TER	0	33
Heterogen	HET	0	34	Threonine	THR	2	7
Histidine	HIS	4	14	Thyroxine	THY	1	23
Homoserine	HSE	2	29	Tryptophan	TRP	1	18
Hydroxyproline	HYP	1	24	Tyrosine	TYR	1	17
Hydroxylysine	HYL	4	22	Valine	VAL	1	3

Notes:- (i) Residue types are:- 1. Hydrophobic 5. Amide
2. Hydrophilic 6. Cyst(e)ine
3. Polar - 7. Glycine
4. Polar + 0. Heterogen

(ii) Residue abbreviations conform to the rules in J. Biol. Chem., 241, 527, 2491 (1966).

(iii) The residue identification numbers have been arbitrarily assigned.

APPENDIX 2

Retrieval Code Numbers and Atom Identifiers

The retrieval code number is a 9-digit number of the form A BB CC DD E.

A is the residue type (hydrophilic etc.) - see Appendix 1.
BBB is the sequence number (field 6 of atomic coordinate record).
CC is the residue identification number - see Appendix 1.
DD is the chain position number - see below.
E is the atom type (field 5 of atomic coordinate record).

In naming the chain position of an atom we use the conventions described in J. Mol. Biol., 52, 1, 1970.

Atom	Identifier	DD	Atom	Identifier	DD
N	N	01	Xε	XE	10
C	C	03	Xε1	XE1	10
O	O	04	Xε2	XE2	11
Cα	CA	02	Xζ	XZ	12
Cβ	CB	05	Xζ1	XZ1	12
Xγ	XG	06	Xζ2	XZ2	13
Xγ1	XG1	06	Xη	XH	14
Xγ2	XG2	07	Xη1	XH1	14
Xδ	XD	08	Xη2	XH2	15
Xδ1	XD1	08			
Xδ2	XD2	09			

Notes:- (i) X represents C, N, O or S

(ii) When nitrogen and oxygen atoms are not distinguished, e.g. NOE1 NOE2 etc., the value of DD is set to 20.

(iii) For tryptophan the numbering scheme is as above as far as XD2. Then we have:-

Atom	Identifier	DD
Nε1	NE1	10
Cε2	CEζ	11
Cε3	CE3	12
Cζ2	CZ2	13
Cζ3	CZ3	14
Cζ2	CH2	15

PROTEIN DATA BANK REQUEST FORM

Appendix 4

1. Name

2. Institution and Postal Address

3. Request

Please supply me with data sets for:-

All available proteins in the Data Bank ☐

Proteins listed below ☐

I guarantee that the data supplied to me are to be used for bona-fide research purposes and not used in any commercial enterprise.

4. Magnetic Tape Specifications

I am sending a magnetic tape under separate cover ☐

I enclose a magnetic tape for the data ☐

Please supply as follows:-

*9-track ☐ *1600 bpi ☐ *EBCDIC ☐ *Labelled ☐

800 bpi ☐ ASCII ☐

7-track ☐ 556 bpi ☐ BCDIC ☐ Unlabelled ☐

* Specification preferred by Cambridge.

5. Data Sets Requested

Name of Protein	Contributor	Data

Brookhaven Crystallographic Computer Program

NAME: SEARCH FILES: 5 TIME: $<100_8$ secs.

CORE FL: $63K_8$

ECS FL: $(16*NATOMS)_8$

Purpose: To read a protein data set from the Protein Data Bank, extract a desired subset, and generate a connectivity table for display programs.

PRØGRAM SEARCH(TAPE1,TAPE10,TAPE11,INPUT,ØUTPUT)

Method: Coordinates of points and radii are input sequentially following the protein data set. SEARCH orthogonalizes and scales to Angstroms the atomic coordinates. It then writes atoms within reach onto an extended core (ECØRE). If the number of atoms extracted exceeds a maximum (currently 400), the radius is decreased, the atoms then within reach are counted, and this process is continued until the number of atoms is less than the working maximum. Then, a modification of UMZUG sets up the connectivity table. The total, extracted list is written on file TAPE11.

Internal Operation: SEARCH reads atomic coordinates and writes 8 words per atom into ECØRE

X(1), X(2), X(3), = fractional coord
X(4) = NATØ(NAT) atomic number
X(5) = atom name (coded)
X(6) = AA tags
X(7), X(8) = Encoded values of X(5), X(6)

Using various criteria,

SEARCH reads the above data from ECØRE and selects certain atoms: this list is added to the one above in ECØRE and consists of 8 words per atom:

W(1), W(2), W(3) = fractional coord
W(4), W(5), W(6) = orthogonal Å coord
W(7) = radius from pt.
W(8) = orig. atom input sequence number.

When the final radius is chosen,

UMZUG reads the ECØRE and stores the X,Y,Z (fract.), RX,RY,RZ (orthogonal), MUNCH (1 to max) = input seq. no., W(8).

Input:

One card for each SEARCH--FØRMAT(2I5,2F5.2,3R10)MANUS,NSHIFT,RADIUS,RADMIN,M1,M2,M

MANUS = -1 Terminator
Cols 1-5 0 SEARCH for all atoms within the radius of point: X,Y,Z (see below
1 find the first atom in the list matching M1, then locate all atoms within the radius
2 for each atom matching M1, find all atoms within the radius
3 for each atom matching M1, find all atoms matching M2 and M3 within the radius

NSHIFT = 1 SEARCH asymmetric unit only
Cols 6-10 2 SEARCH symmetrically related atoms also
3 SEARCH symmetrically related atoms and apply translations to the 26 neighboring cells

11-15 RADIUS maximum radius
16-20 RADMIN minimum radius
22-30 M1
32-40 M2 or {X, Y, Z} (coordinates of origin point of SEARCH decoded into 3F10.5 format)
42-50 M3 IF MANUS=0

M1, M2, M3 are 9-digit SEARCH codes in the form A/BBB/CC/DD/E as defined by PRØIN:

A = aqueous property
1 = hydrophobic
2 = hydrophyllic
3 = polar, -charge
4 = polar, +charge
5 = amide residue
6 = cyst (ei/i)ne
7 = glycine

B = amino acid sequence number
C = amino acid type (cf. dictionary)
D = numerical expression of the position of the atom of Appendix 2, Protein Data Bank File Record Format
E = atom type, according to the order of the Protein Data File

TAPE1 contains a protein file from the Protein Data Bank. This may be gotten from the data bank UPDATE tape.

Output:

The job's output file contains cell and symmetry information, selected atoms, connectivities, and a copy of TAPE1.

TAPE10: comments and error messages

TAPE11: selected atomic coordinates
Card 1. (5I3,2X,5A10)
cols 1-5 IFRØM number of atoms selected
cols 4-6 IRING number of ring closures for connectivity
cols 7-9 NCENT =1 for acentric
=2 for centric
cols 10-12 NØRTH not used
cols 13-15 NEQV number of symmetry cards
cols 18-68 TITLE from CØMPND card in data bank

Card 2. (6F9.3) ABC,ANG
cols 1-9 a
cols 10-18 b
cols 19-27 c
cols 28-36 α degrees
cols 37-45 β
cols 46-54 γ

Cards 3. (A10,17X,3F9.6,8X,A10) compatible with FLINUS
cols 1-10 DU(6) atom designation (amino acid name in cols 1-3; residue number in cols 4-6; CA, CB, etc in cols 7-10
cols 28-36 X
cols 37-45 Y
cols 46-54 Z
cols 63-72 M (9 digit code)

Cards 4. (15I5)
connectivity "from" list IFRØM total entries

Cards 5. (15I5)
Ring closure list IRING total entries

Cards 6. (3(F15.0,3F3.0)) NEQV Symmetry cards
FLINUS format

Card 7. (I5)
cols 4-5 -0 terminator

Several examples of SEARCH procedures may be helpful:

Example 1. The active site contains His 57; extract all atoms within 10.0 Å of the epsilon 2 Nitrogen of His 57
M1 = 0/057/14/11/2
the input file would then contain one data card as follows:
1 1 10.0 005714112

Example 2. Find all charged groups within 7.0 Angstroms of one of the terminal N atoms in lysine in a protein known to have no cysteine.
M1 = 0/000/15/14/0
M2 = 0/000/00/00/2 Nitrogen atom
M3 = 0/000/00/00/3 Oxygen atoms
the data card would be
3 3 7.0 000015140 000000002 000000003
this will find all oxygen or nitrogen atoms within 7 Å of the terminal N atom of each lysine, with symmetry and translation operations added.

Example 3. Locate all O-N or O-O hydrogen bonds within a given molecule
M1 = 0/000/00/00/3 3 = oxygen
M2 = 0/000/00/00/3
M3 = 0/000/00/00/2 2 = nitrogen
the data card would be:
3 1 3.5 2.5 000000003 000000003 000000002
locates all oxygen atoms in the structure and then SEARCHES for all N or O atoms within 2.5 to 3.5 Å (assumed max. H bonding distance)

SAMPLE DECK SETUP

```
ØBTAIN(CPL,CPRØGS,PDB,ØLDPL)  get Data Bank update file
BUPD140(Q,C=TAPE1)
RETURN(ØLDPL)
ØBTAIN(CPRØGS,CPRØGS,SEARCH)
SEARCH.
REWIND(TAPE10,TAPE11)
CØPYSBF(TAPE10,ØUTPUT)
CØPYSBF(TAPE11,ØUTPUT)   repeat as needed for each SEARCH performed
7/8/9
*C,DECK       deck is the name of the desired protein on the update tape
7/8/9
[search data cards]
6/7/8/9
```

PRØTEIN DATA BANK

DECK NAMES FOR PRØTEINS IN THE MASTER UPDATE FILE

DECK NAME	PRØTEIN
CHYMØ1	α-CHYMØTRYPSIN
CPASE01	CARBØXYPEPTIDASE
CYTB501	CYTØCHRØME B5
HSDEHB01	HØRSE DEØXYHEMØGLØBIN
HSMEHB01	HØRSE METHEMØGLØBIN
INSULO1	INSULIN
LDHO1	LACTATE DEHYDRØGENASE (SET 1)
LDHO2	LACTATE DEHYDRØGENASE (SET 2)
LAMP1	LAMPREY HEMØGLØBIN
MYØGLO1	MYØGLØBIN
PAPAINO1	PAPAIN
RNASESO1	RIBØNUCLEASE-S
RUBYO1	RUBREDØXIN
STAPHNO1	STAPH NUCLEASE
SUBTLE1	SUBTILISIN
TRINO1	PANCREATIC TRYPSIN INHIBITØR (SET 1)
TRINO2	PANCREATIC TRYPSIN INHIBITØR (SET 2)