PROTEIN DATA BANK

FILE RECORD FORMATS

(140 Column Version)

SEPTEMBER 1972

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.

RECORD FORMATS

1. COMPND	Cols.	1-6	COMPND	(6A1)
		11-70	Name of Protein	(60A1)
		73-80	Identification Code	(8A1)
2. AUTHOR	Cols.	1-6	AUTHOR	(6A1)
		11-70	Names of Contributor and Co-Authors	(60A1)
		73-80	Identification Code	(8A1)
3. CRYST1	Cols.	1-6	CRYST1	(6A1)
		8-15	a(Å)	(F8.3)
		17-24	b(Å)	(F8.3)
		26-33	c(Å)	(F8.3)
		35-40	α(deg.)	(F6.2
		42-47	β(deg.)	(F6.2)
		49-54	γ(deg.)	(F6.2)
		56-66	Space Group Symbol	(11A1)
		68-70	Z	(I3)
		73-80	Identification Code	(8A1)
4. DECODE	Cols.	1-6	DECODE	(6A1)
		11-70	At. Number ; Element Symbol Pairs	(10(I2,1X,A3))
		73-80	Identification Code	(8A1)

Note:- As an example, the data in cols. 11-70 might be:-

6C 7N 8O 16S

5.	REMARK	Cols.	1-6	REMARK (6A1)
			8-9	Serial Number (I2)
			11-70	Remarks (60A1)
			73-80	Identification Code (8A1)

Note:- The first REMARK card has serial number 1, the second 2 etc. etc.

6.	FTNOTE	Cols.	1-6	FTNOTE (6A1)
			8-9	Footnote Number (I2)
			11-70	Footnote Statement (60A1)
			73-80	Identification Code (8A1)

Note:- FTNOTE records are used to describe details which are specific to certain atoms or residues.

e.g. One might wish to indicate that certain atomic co-ordinates are very unreliable.
Suppose this is the first of a series of footnotes.
Then we have 1 in col. 9 and the descriptive statement in cols. 11-70. The atomic coordinate records for these atoms would carry 1 in col. 132, thus linking them to footnote number 1.

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 $\mathring{A}$ 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 $\underset{\sim}{a} \times \underset{\sim}{b}$.

Note that in the new (80 column) file the SCALE and ORIGX transformations are defined differently from those above.

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	CE2	11
Cε3	CE3	12
Cζ2	CZ2	13
Cζ3	CZ3	14
Cζ2	CH2	15

UNIVERSITY CHEMICAL LABORATORY,
LENSFIELD ROAD,
CAMBRIDGE,
CB2 1EW
TELEPHONE (0223) 56491

Dear Colleague,

PROTEIN DATA BANK

You may be interested to know that the repository system for protein crystallographic data (Nature New Biology, 233, 223, 1971) is now in operation. To facilitate the deposition and retrieval of data we have prepared the following documents which are enclosed with this letter:-

(1) List of proteins already on file
(2) File record formats and specimen listing
(3) Input form
(4) Data request form.

We are able to accept data in any specified format. It would, however, greatly facilitate our task if magnetic tapes are written as follows:- 7-track, 556 bpi, unlabelled, unblocked.

Submitted data will be organised according to our format specifications and a listing sent to you for your approval before the data set is filed and made available for distribution.

It would be of great assistance to us to receive copies of any relevant preprints or reprints.

The data request forms are to be used for obtaining data from the Protein Data Bank.

Identical files are being maintained in England and the U.S.A. and data can be submitted to, and obtained from, any of the addresses listed below.

Yours sincerely,

Edgar Meyer,
Texas A & M University,
College of Agriculture,
Dept. of Biochemistry & Biophysics,
College Station, Texas 77843.

Olga Kennard, David Watson,
University Chemical Laboratory,
Lensfield Road,
Cambridge, England, CB2 1EW.

Walter Hamilton, Helen Berman,
Dept. of Chemistry,
Brookhaven National Laboratory,
Upton, L.I., New York 11973.

UNIVERSITY CHEMICAL LABORATORY,
LENSFIELD ROAD,
CAMBRIDGE,
CB2 1EW
TELEPHONE (0223) 56491

Dear Colleague,

PROTEIN DATA BANK

Thank you very much for your offer to deposit data in the Protein Data Bank. To facilitate the deposition and retrieval of data we have prepared the following documents which are enclosed with this letter:-

(1) Lists of proteins already on file
(2) File record formats and specimen listing
(3) Input form
(4) Data request form.

We are able to accept data in any specified format. It would, however, greatly facilitate our task if magnetic tapes are written as follows:- 7-track, 556 bpi, unlabelled, unblocked.

Submitted data will be organised according to our format specifications and a listing sent to you for your approval before the data set is filed and made available for distribution.

It would be of great assistance to us to receive copies of any relevant preprints or reprints.

The data request forms are to be used for obtaining data from the Protein Data Bank.

Identical files are being maintained in England and the U.S.A. and data can be submitted to, and obtained from, any of the addresses listed below.

Yours sincerely,

Edgar Meyer,
Texas A & M University,
College of Agriculture,
Dept. of Biochemistry & Biophysics,
College Station, Texas 77843.

Olga Kennard, David Watson
University Chemical Laboratory,
Lensfield Road,
Cambridge, England, CB2 1EW.

Walter Hamilton, Helen Berman,
Dept. of Chemistry,
Brookhaven National Laboratory,
Upton, L.I., New York 11973.

PROTEIN DATA BANK REQUEST FORM

Please supply me with the following data sets:-

Name of Protein	AC	TA	SF	ED

AC: Atomic Coordinates
TA; Torsion Angles
SF: Structure Factors
ED: Electron Densities

I enclose a magnetic tape for this service.

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

Signed:

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

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)