PROTEIN DATA BANK

FILE RECORD FORMATS

February 1978

Contents

MACROMOLECULAR STRUCTURE FILE ENTRY

For each parameter set the file consists of records each of 80 characters. The record sequence is as follows:

HEADER	:	Date entered into data bank; identification code
COMPND	:	Name of molecule and identifying information
SOURCE	:	Species from which molecule has been obtained
AUTHOR	:	Names of contributor and coauthors
JRNL	:	Literature citation that defines coordinate set
REMARK	:	General remarks
SEQRES	:	The residue sequence
FTNOTE	:	Footnotes relating to specific atoms or residues
HET	:	Identification of non-standard groups or residues (heterogens)
FORMUL	:	Chemical formulas of non-standard groups
HELIX	:	Identification of helical substructures
SHEET	:	Identification of sheet substructures
TURN	:	Identification of hairpin turns
SSBOND	:	Specification of disulfide bonds
SITE	:	Identification of groups comprising the various sites
CRYST1	:	Unit cell parameters, space group designation
ORIGX	:	Transformation from orthogonal Å coordinates to submitted coordinates
SCALE	:	Transformation from orthogonal Å coordinates to fractional crystallographic coordinates
MTRIX	:	Transformations expressing non-crystallographic symmetry
ATOM	:	Atomic coordinate records
SIGATM	:	Standard deviations in atomic parameters

HETATM : Coordinates for non-standard groups

TER : Chain terminator

CONECT : Connectivity records

MASTER : Master control record with checksums of total number of records in the file, broken down by record type

END : End-of-entry record

In describing record formats it will be convenient to use the punched-card analogy and refer to column numbers. Although up to six characters are used in the record tag words above, only the first four are needed to define the record type uniquely.

RECORD FORMATS

1. HEADER Cols. 1-6 HEADER

   11-50 Functional classification of macromolecule

   51-60 Date entered into data bank

   63-66 Identification code

   FORMAT (6A1,4X,50A1,2X,4A1)

   Note: Each macromolecule is assigned an identification code. The code consists of 1 numeric and 3 alphanumeric characters. The former is provided to distinguish multiple data sets for the same macromolecule.

2. COMPND Cols. 1-6 COMPND

   11-60 Name of molecule (The macromolecule will be identified by both trivial and systematic names where appropriate).

   FORMAT (6A1,4X,50A1)

   Note: For enzymes the E.C. number is given in the form (E.C.n.n.n.n.) with no internal blanks and without splitting over two lines. If an enzyme has not had an E.C. number assigned the string (E.C. NUMBER NOT ASSIGNED) will be used.

3. SOURCE Cols. 1-6 SOURCE

   11-60 Species, tissue, mutant from which the molecule has been obtained. The systematic name of the species is given in parentheses.

   FORMAT (6A1,4X,50A1)

4. AUTHOR Cols. 1-6 AUTHOR

   11-60 Name(s) of contributor(s)

   FORMAT (6A1,4X,50A1)

5. JRNL Cols. 1-4 JRNL

   11-70 Literature citation that defines the coordinate set

   FORMAT (6A1,4X,60A1)

NOTE: See appendices E and F for detailed specifications.

6. <u>REMARK</u> Cols. 1-6 REMARK

8-10 Remark number

12-70 Text of remark

FORMAT (6A1,1X,I3,60A1)

<u>Note</u>: - The first REMARK has serial number 1, the second 2, etc. See appendices E and F for detailed specifications.

7. <u>SEQRES</u> Cols. 1-6 SEQRES

9-10 Serial number

12 Chain identifier

14-17 Number of residues in this chain

20-22 Residue name

24-26 Residue name

.

.

.

68-70 Residue name

FORMAT (6A1,I4,1X,A1,1X,I4,1X,13(1X,A3))

<u>Note</u>: Serial numbers are reset to 1 for each new chain.

8. <u>FTNOTE</u> Cols. 1-6 FTNOTE

8-10 Footnote number

12-70 Footnote statement

FORMAT (6A1,1X,I3,60A1)

<u>Note</u>: - FTNOTE records are used to describe details which are specific to certain atoms or residues. These footnotes are keyed to particular atoms by the footnote number here and in cols 68-70 of the ATOM record. Any individual footnote may run over several FTNOTE records (each with the same footnote number). A maximum of 999 FTNOTEs are allowed.

9. HET Cols. 1-3 HET

8-10 Non-standard group (heterogen) identifier

13 Chain identifier

14-17 Sequence number

18 Insertion code

21-25 Number of atoms in non-standard group

31-70 Text

FORMAT (6A1,1X,A3,2X,A1,I4,A1,2X,I5,5X,40A1)

Note: - HET records are used to describe non-standard residues, prosthetic groups, inhibitors, solvent molecules (except water) etc. All non-standard components (i.e., those not assigned a standard code in appendix D) are defined in these records. If there is insufficient space in the text portion of the HET record to properly define a non-standard component then the definition will be given in a REMARK and referenced here.

10. FORMUL Cols. 1-6 FORMUL

9-10 Component number (i)

13-15 Non-standard group (HET) identifier

17-18 Continuation number (ii) (blank on first record)

19 * if this component is to be excluded from the molecular weight calculation. (iii)

20-70 Formula of non-standard group (iv)

FORMAT (6A1,2X,I2,2X,A3,1X,I2,A1,51A1)

Notes: (i) Component numbers are assigned serially. Each component represented by a set of SEQRES records is counted first and then each HET group is assigned a component number in sequence. If a HET group is contained within a chain represented by a set of SEQRES records (e.g. the "Y" base of the tRNA's) the component number assigned is that of the chain involved.

(ii) If a HET group is composed of more than one distinct part, then the formulas for these parts will occur on separate FORMUL cards each with the same component number and HET identifier. All except the last of these records will be terminated with a period.

(iii) Solvent molecules and certain other components are normally excluded. The molecular weight is used as a key for automatic searching of the file.

(iv) Each component defined in a HET record for which a standard chemical formula can be written is defined accordingly here. Atoms which are known to be present but not located in the crystallographic analysis (e.g. hydrogen atoms) are represented in the formula. Formulas are written as C,H,N,O with other elements following in alphabetical order. The repeat count of each atom type present immediately follows the chemical symbol. A repeat count of the entire group is indicated by enclosing the formula in parentheses and prefacing the string with the count. The ionization state of metals is given when it is known. For the two heme groups of Ferrihemoglobin the FORMUL record would have HEM 2(C34 H32 N4 O4 FE1 +++)

11. <u>HELIX</u> Cols.

1-6	HELIX	
8-10	Serial number (Helix number)	
12-14	Helix identifier (right justified)(i)	
16-18	Residue name	Initial residue of helix(ii)
20	Chain identifier	
22-25	Residue seq. no.	
26	Insertion code	
28-30	Residue name	Terminal residue of helix
32	Chain identifier	
34-37	Residue seq. no.	
38	Insertion code	

39-40 Class of helix(iii)

41-70 Comment

FORMAT (6A1,1X,I3,1X,A3,2(1X,A3,1X,A1,1X,I4,A1),I2,30A1)

Notes: (i) Additional records with different serial numbers and identifiers occur if more than one helix is present.

(ii) The initial residue has a lower sequence number than the terminal resolve.

(iii) Helices are classified as:

1	Right-handed α (default)	2	Right-handed ω
3	Right-handed π	4	Right-handed γ
5	Right-handed 3_{10}	6	Left-handed α
7	Left-handed ω	8	Left-handed γ
9	2_7 ribbon/helix		

12. <u>SHEET</u> Cols. 1-5 SHEET

8-10 Strand number(i)(v)

12-14 Sheet identifier(i) (right justified)

15-16 Number of strands

18-20 Residue name

22 Chain identifier

23-26 Residue seq. no. Initial residue(ii)

27 Insertion code

29-31 Residue name

33 Chain identifier

34-37 Residue seq. no. Terminal residue

38 Insertion code

39-40 Sense of this strand with respect to previous strand(iii)

42-45	Atom name	
46-48	Residue name	
50	Chain identifier	Registration (iv)
51-54	Residue seq. no.	
55	Insertion code	
57-60	Atom name	
61-63	Residue name	
65	Chain identifier	Registration (iv)
66-69	Residue seq. no.	
70	Insertion code	

FORMAT (6A1,I4,1X,A3,I2,2(1X,A3,1X,A1,I4,A1),I2,
2(1X,A4,A3,1X,A1,I4,A1))

Notes:

(i) Different strands are described in subsequent records which bear the same sheet identifier but different strand numbers.

(ii) The initial residue of a strand has a lower sequence number than the terminal residue.

(iii) Parallelism or anti-parallelism of strand n with respect to strand n-1 is denoted by 1 or -1. Strand 1 has sense indicator 0.

(iv) Registration of the strand n with respect to strand n-1 may be specified by a particular hydrogen bond between the indicated atoms. One donor and one acceptor should be specified. These fields will be blank for strand 1.

(v) Strand numbers are reset to 1 for the first strand of each new sheet. A closed sheet (β barrel) is indicated by having the first and last strands identical.

13. TURN Cols. 1-5 TURN

8-10	Sequence number (Turn number)
12-14	Turn identifier (3 characters)

16-18	Residue name	
20	Chain identifier	Residue i
21-24	Residue seq. no.	
25	Insertion code	
27-29	Residue name	
31	Chain identifier	Residue i + 3 (or i+2 for γ bend)
32-35	Residue seq. no.	
36	Insertion code	
41-70	Comment	

FORMAT(6A1,1X,I3,1X,A3,1X,A3,1X,A1,I4,1X,A1,A3,1X,A1,I4,A1,4X,30A1)

Note: - These records identify the hairpin turns (β and γ bends) in the structure which do not occur in helices.

14. SSBOND Cols. 1-6 SSBOND

8-10	Sequence number
12-14	Residue name (CYS)
16	Chain identifier
18-21	Residue seq. no.
22	Insertion code
26-28	Residue name (CYS)
30	Chain identifier
32-35	Residue sequence number
36	Insertion code
41-70	Comment

FORMAT (6A1,1X,I3,1X,A3,1X,A1,1X,I4,A1,3X,A3,1X,A1,1X,I4,A1,4X,30A1)

15. <u>SITE</u>

Cols. 1-4	SITE	
8-10	Sequence number[i]	
12-14	Site identifier[ii]	(right justified)
16-17	Number of residues comprising site[iii]	
19-21	Residue name	First residue comprising site
23	Chain identifier	
24-27	Residue seq. no.	
28	Insertion code	
30-32	Residue name	Second residue comprising site
34	Chain identifier	
35-38	Residue seq. no.	
39	Insertion code	
41-43	Residue name	Third residue comprising site
45	Chain identifier	
46-49	Residue seq. no.	
50	Insertion code	
52-54	Residue name	Fourth residue comprising site
56	Chain identifier	
57-60	Residue seq. no.	
61	Insertion code	

FORMAT (6A1,1X,I3,1X,A3,1X,I2,4(1X,A3,1X,A1,I4,A1))

Notes: (i) Sequence numbers are reset to 1 for each new site.

(ii) Site identifiers should be fully explained in the REMARKs.

(iii) If a site is comprised of more than four residues then these may be specified on additional records bearing the same site identifier.

16. CRYST1 Cols. 1-6 CRYST1

Cols.	
1-6	CRYST1
7-15	a(Å)
16-24	b(Å)
25-33	c(Å)
34-40	α(deg.)
41-47	β(deg.)
48-54	γ(deg.)
56-66	Space group symbol (left justified)
67-70	z

FORMAT (6A1,3F9.3,3F7.2,1X,11A1,I4)

17. ORIGX

Cols. 1-6	11-20	21-30	31-40	46-55
ORIGX1	O_{11}	O_{12}	O_{13}	T_1
ORIGX2	O_{21}	O_{22}	O_{23}	T_2
ORIGX3	O_{31}	O_{32}	O_{33}	T_3

FORMAT (6A1,4X,3F10.5,5X,F10.5)

Note: - Let the original deposited coordinates be X_o, Y_o, Z_o.

Let the orthogonal Å coordinates contained in the file be X,Y,Z. Then

$$X_o = O_{11}X + O_{12}Y + O_{13}Z + T_1$$

$$Y_o = O_{21}X + O_{22}Y + O_{23}Z + T_2$$

$$Z_o = O_{31}X + O_{32}Y + O_{33}Z + T_3$$

Even if this is an identity transformation (unit matrix, null vector) it is supplied. See below under SCALE for a definition of the default orthogonal Å system.

Appendix A details the derivation of this coordinate transformation.

18. <u>SCALE</u> Cols 1-6 11-20 21-30 31-40 46-55

SCALE1	S_{11}	S_{12}	S_{13}	U_1
SCALE2	S_{21}	S_{22}	S_{23}	U_2
SCALE3	S_{31}	S_{32}	S_{33}	U_3

FORMAT (6A1,4X,3F10.5,5X,F10.5)

Note: - Let the orthogonal Å coordinates be X,Y,Z. Let the fractional cell coordinates be X_f, Y_f, Z_f. Then

$$X_f = S_{11}X + S_{12}Y + S_{13}Z + U_1$$

$$Y_f = S_{21}X + S_{22}Y + S_{23}Z + U_2$$

$$Z_f = S_{31}X + S_{32}Y + S_{33}Z + U_3$$

The SCALE transformation provides a means of generating fractional coordinates from the orthogonal Å coordinates contained in the file.

Unless otherwise specified in the REMARKs the orthogonal Å coordinate system is related to the axial system of the space group supplied (CRYST1 record) by the definition below.

If $\vec{A},\vec{B},\vec{C}$ are unit vectors in the orthogonal Å system

and $\vec{a},\vec{b},\vec{c}$ are unit vectors in the crystallographic system, then:

(i) the $\vec{A},\vec{B},\vec{C}$ and $\vec{a},\vec{b},\vec{c}$ systems have the same origin

(ii) $\vec{A}$ is parallel to $\vec{a}$

(iii) $\vec{B}$ is parallel to $\vec{C} \times \vec{A}$

(iv) $\vec{C}$ is parallel to $\vec{a} \times \vec{b}$, (i.e. $\vec{c}^*$)

Appendix A details the derivation of this coordinate transformation.

19. MTRIX

Cols. 1-6	8-10	11-20	21-30	31-40	46-55
MTRIX1	Ser.no.	M_{11}	M_{12}	M_{13}	V_1
MTRIX2	Ser.no.	M_{21}	M_{22}	M_{23}	V_2
MTRIX3	Ser.no.	M_{31}	M_{32}	M_{33}	V_3

FORMAT (6A1,1X,I3,3F10.5,5X,F10.5)

Note: - In a structure with non-crystallographic symmetry, MTRIX records are used to generate coordinates for the full contents of an asymmetric unit. In some cases the full asymmetric unit is explicitly given and in these cases the MTRIX records simply describe the nature of the local symmetry operations.

Let orthogonal Å coordinates of a subunit contained in the file be X, Y, Z and coordinates of a particular related subunit be X', Y', Z'. Then

$$X' = M_{11}X + M_{12}Y + M_{13}Z + V_1$$

$$Y' = M_{21}X + M_{22}Y + M_{23}Z + V_2$$

$$Z' = M_{31}X + M_{32}Y + M_{33}Z + V_3$$

The serial number is constant for each group of three MTRIX records. More than one group of MTRIX records may be present to describe different local symmetry operations. Appendix A details the derivation of this coordinate transformation.

20. ATOM Atomic coordinate records

Cols.	
1-4	ATOM
7-11	Atom serial number(i)
13-16	Atom name(ii)
17	Alternate location indicator(iii)
18-20	Residue name(iv)
22	Chain identifier, e.g., A for hemoglobin α chain
23-26	Residue sequence number
27	Code for insertions of residues, e.g., 66A, 66B, etc.

31-38	X	
39-46	Y	Orthogonal Å coordinates
47-54	Z	
55-60	Occupancy	
61-66	Temperature factor	
68-70	Footnote number	

FORMAT (6A1,I5,1X,A4,A1,A3,1X,A1,I4,A1,3X,3F8.3,2F6.2,1X,I3)

Note: - (i) Residues occur in order of their sequence numbers which always increase starting from the N-terminal residue for proteins and 5'-terminal for nucleic acids. Within each residue the atoms are ordered as indicated in Appendix B. If the residue sequence is known certain atom serial numbers may be omitted to allow for future insertion of any missing atoms. If the sequence is not reliably known these serial numbers are simply ordinals.

(ii) Appendix B

(iii) Alternate locations for atoms may be denoted by A, B, C, etc. here.

(iv) Standard residue names are given in Appendix C; other components are defined in HET records.

21. SIGATM Cols. 1-6 SIGATM

7-27	Identical to corresponding ATOM record
31-38	
39-46	Standard deviations in the stored coordinates (Å)
47-54	
55-60	Standard deviation in occupancy
61-66	Standard deviation in temperature factor
68-70	Footnote number

FORMAT (6A1,I5,1X,A4,A1,A3,1X,A1,I4,A1,3X,3F8.4,2F6.3,1X,I3)

22. HETATM Cols. 1-6 HETATM

7-70 As for ATOM records

Note: - HETATM records are used for water molecules and atoms contained in HET groups.

23. TER Cols. 1-3 TER

7-11 Serial number

18-20 Residue name

22 Chain identifier

23-26 Residue sequence number

27 Insertion code

FORMAT (6A1,I5,6X,A3,1X,A1,I4,A1)

Note: - TER records occur among the ATOM records, and are placed after the terminal atom of each chain. For a protein the residue defined on these TER records is the carboxy-terminal residue of the chain in question. For a nucleic acid it is the 3'-terminal residue.

24. CONECT Connectivity records

Cols.		
1-6	CONECT	
7-11	Serial number	
12-16	Covalent bond connectivity (serial numbers of bonded atoms)	
17-21		
22-26		
27-31		
32-36	Hydrogen bond	in which the atom specified in cols. 7-11 acts as donor
37-41	Hydrogen bond	
42-46	Salt bridge	the atom specified in cols. 7-11 has an excess of negative charge

47-51	Hydrogen bond	in which the atom specified in cols. 7-11 acts as acceptor
52-56	Hydrogen bond	
57-61	Salt bridge	the atom specified in cols. 7-11 has an excess of positive charge

FORMAT (6A1,11I5)

Note: - Serial numbers are identical to those in cols. 7-11 of the appropriate ATOM/HETATM records and connectivity entries correspond to these serial numbers. A second CONECT record, with the same serial number in cols. 7-11, may be used if necessary. Either all or none of the covalent connectivity of an atom must be specified, and if hydrogen bonding is specified the covalent connectivity is included also.

25. MASTER Cols. 1-6 MASTER

11-15	Number of REMARK records
16-20	Number of FTNOTE records
21-25	Number of HET records
26-30	Number of HELIX records
31-35	Number of SHEET records
36-40	Number of TURN records
41-45	Number of SITE records
46-50	Number of coordinate transformation records (ORIGX+SCALE+MTRIX)
51-55	Number of atomic coordinate records (ATOM+HETATM)
56-60	Number of TER records
61-65	Number of CONECT records
66-70	Number of SEQRES records

FORMAT (6A1,4X,12I5)

Note: - The MASTER record gives checksums of the number of records in the file, broken down by record type.

26. <u>END</u> End-of-entry record

Cols. 1-3 END

FORMAT (6A1)

APPENDIX A

Coordinate Transformations

In the equations of this appendix matrices and vectors are denoted by upper and lower case characters, respectively.

Usually the depositor will have supplied--

(i) The original submitted coordinates, i.e. x_o

(ii) A transformation from x_o to the orthogonal Å coordinates stored in the data bank (x), i.e.,

$$O_o x_o + t_o = x$$

where the subscripts denote "in the space of the original coordinates".

The O_o, t_o transformation which transforms fractional crystallographic coordinates into the default orthogonal Å system is

$$O_o = \begin{matrix} a & b\cos\gamma & c\cos\beta \\ 0 & b\sin\gamma & c(\cos\alpha - \cos\beta\cos\gamma)/\sin\gamma \\ 0 & 0 & V/(ab\sin\gamma) \end{matrix}\quad , \quad t_o = \begin{matrix} 0 \\ 0 \\ 0 \end{matrix}$$

where the unit cell volume is:

$$V = abc(1 - \cos^2\alpha - \cos^2\beta - \cos^2\gamma + 2\cos\alpha\cos\beta\cos\gamma)^{\frac{1}{2}}$$

(iii) A transformation from x_o to fractional crystallographic coordinates (x_f)

i.e. $S_o x_o + u_o = x_f$

(iv) A set of transformations expressing any non-crystallographic symmetry elements in the structure

i.e. $M_o x_o + v_o = x_o'$

Since it is desirable for the stored ORIGX, SCALE and MTRIX transformations to operate on the stored rather than the submitted coordinates some manipulation of the supplied quantities is performed in order to obtain the stored quantities.

The stored quantities are:

(i) The coordinates in orthogonal Ångstroms (x)

$$x = O_o x_o + t_o$$

(ii) The ORIGX transformation from stored to original coordinates (O,t)

From above $x = O_o x_o + t_o$

whence $O_o x_o = x - t_o$

$$x_o = O_o^{-1} x + (-O_o^{-1} t_o)$$

Thus $O = O_o^{-1}$

and $t = -O^{-1} t_o$

(iii) The SCALE transformation from stored to fractional coordinates (S,u).

From above $x_f = S_o x_o + u_o$

but $x_o = O_o^{-1} x + (-O_o^{-1} t_o)$

$$x_f = S_o\left(O_o^{-1} x + (-O_o^{-1} t_o)\right) + u_o$$

i.e. $$x_f = S_o O_o^{-1} x + (-S_o O_o^{-1} t_o) + u_o$$

$$S = S_o O_o^{-1}$$

and $$u = -(S_o O_o^{-1} t_o) + u_o$$

(iv) The MTRIX transformation(s) expressing non-crystallographic symmetry in the space of the stored co-ordinates (M,v).

$$x_o' = M_o x_o + v_o$$

$$x' = O_o x_o' + t_o$$

$$= O_o (M_o x_o + v_o) + t_o$$

but $$x_o = O_o^{-1} x + (-O_o^{-1} t_o)$$

and so $$x' = O_o M_o \left(O_o^{-1} x + (-O_o^{-1} t_o)\right) + v_o + t_o$$

whence $$M = O_o M_o O_o^{-1}$$

and $$v = -O_o M_o O_o^{-1} t_o + O_o v_o + t_o$$

In summary the stored coordinates and transformations are:

X	(ATOM records)
O,T	(ORIGX records)
S,U	(SCALE records)
M,V	(MTRIX records)

APPENDIX B

Atom Identifiers

A. Amino Acids

These atom names follow the IUPAC-IUB rules[1] except:

(i) Greek letter remoteness codes are transliterated according to the following table

α-A	δ-D	η-H
β-B	ε-E	
γ-G	ζ-Z	

(ii) Atoms for which some ambiguity exists in the crystallographic results are designated A. This will usually apply only to the terminal atoms of asparagine and glutamine and to the ring atoms of histidine.

Within each residue the atoms occur in the order specified by the superscripts (following figure).

The extra oxygen atom of the carboxy terminal amino acid is designated OXT.

Four characters are reserved for these atom names--they are assigned as follows:

1-2 Chemical symbol - right justified

3 Remoteness indicator (alphabetic)

4 Branch designator (numeric)

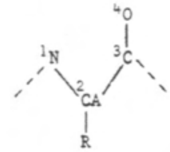
---

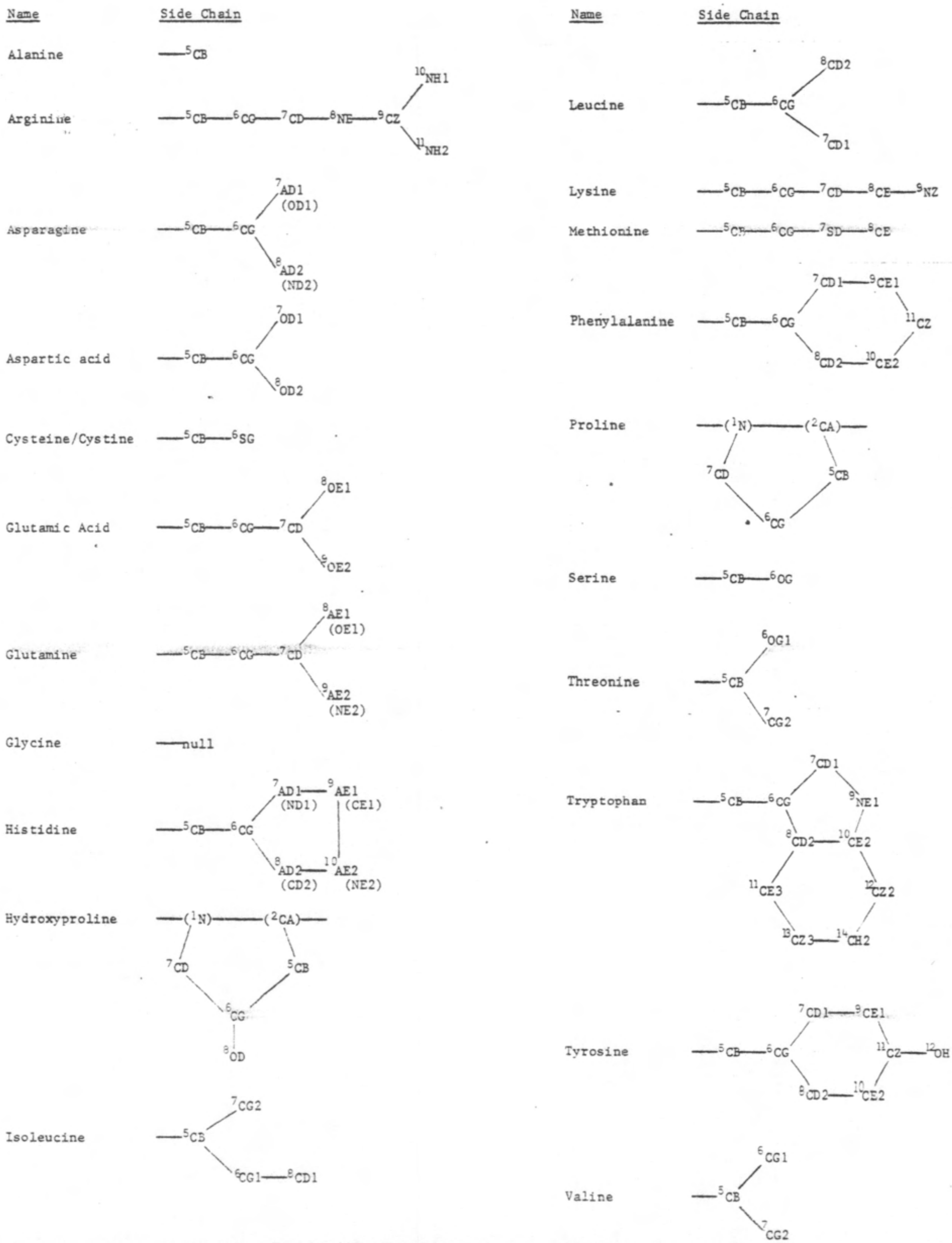
[1]IUPAC-IUB Commission on Biochemical Nomenclature. "Abbreviations and Symbols for the Description of the Conformation of Polypeptide Chains. Tentative Rules (1969)," J. Biol. Chem 245, 6489 (1970).

The 1974 recommendations on the "Nomenclature of α-Amino Acids (Biochemistry, 14, 449 (1975)) provides a scheme based on normal rules for organic compounds but this scheme will not be used here.

Atom Names, Remoteness Codes, and Order Indicators for the Common Amino Acids

backbone





B. Nucleic Acids

Atom names employed for polynucleotides generally follow the precedents set for mononucleotides. The following points are worthy of note.

(i) The prime character (') commonly used to denote atoms of the ribose is eschewed because of non-uniformity of its external representation. An asterisk (*) is used in its place.

(ii) Of the four characters reserved for atom names the leftmost two are reserved for the chemical symbol (right justified), the remaining two denote the atom's position.

(iii) Atoms exocyclic to the ring systems have the same position identifier as the atom to which they are bonded except if identical atom names would result. In this case an alphabetic character is used to avoid ambiguity.

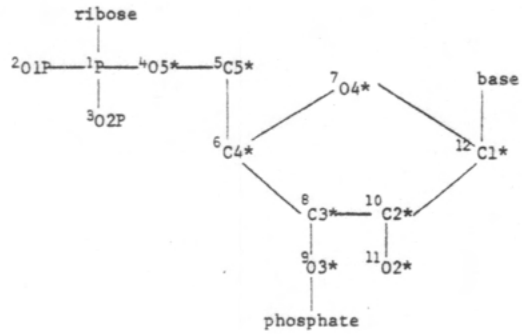
(iv) The ring-oxygen atom of the ribose is denoted O4 and not O1.

(v) The extra oxygen atom at the free 5' phosphate terminus is designated OXT. This atom will be placed first in the coordinate set.

For nucleotides which are simple derivatives (e.g. methyl or acetyl) of the parent nucleotide the modifying atoms or groups occur immediately after the atom to which they are bonded. In the case of an acetyl modifier the three atoms are ordered carbonyl carbon, carbonyl oxygen, methyl carbon.

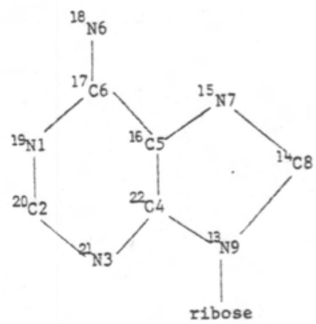
Atom names and order indicators for the common ribonucleotides (superscripts denote ordinals).

backbone

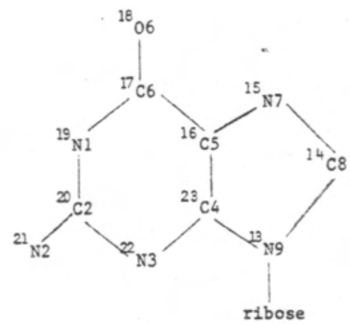


bases (names according to nucleoside)

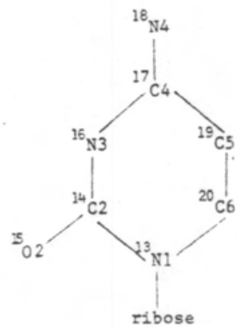
Adenosine



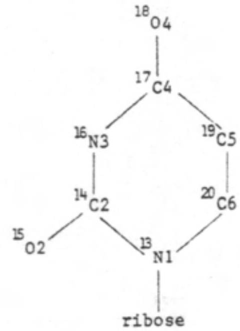
Guanosine



Cytidine

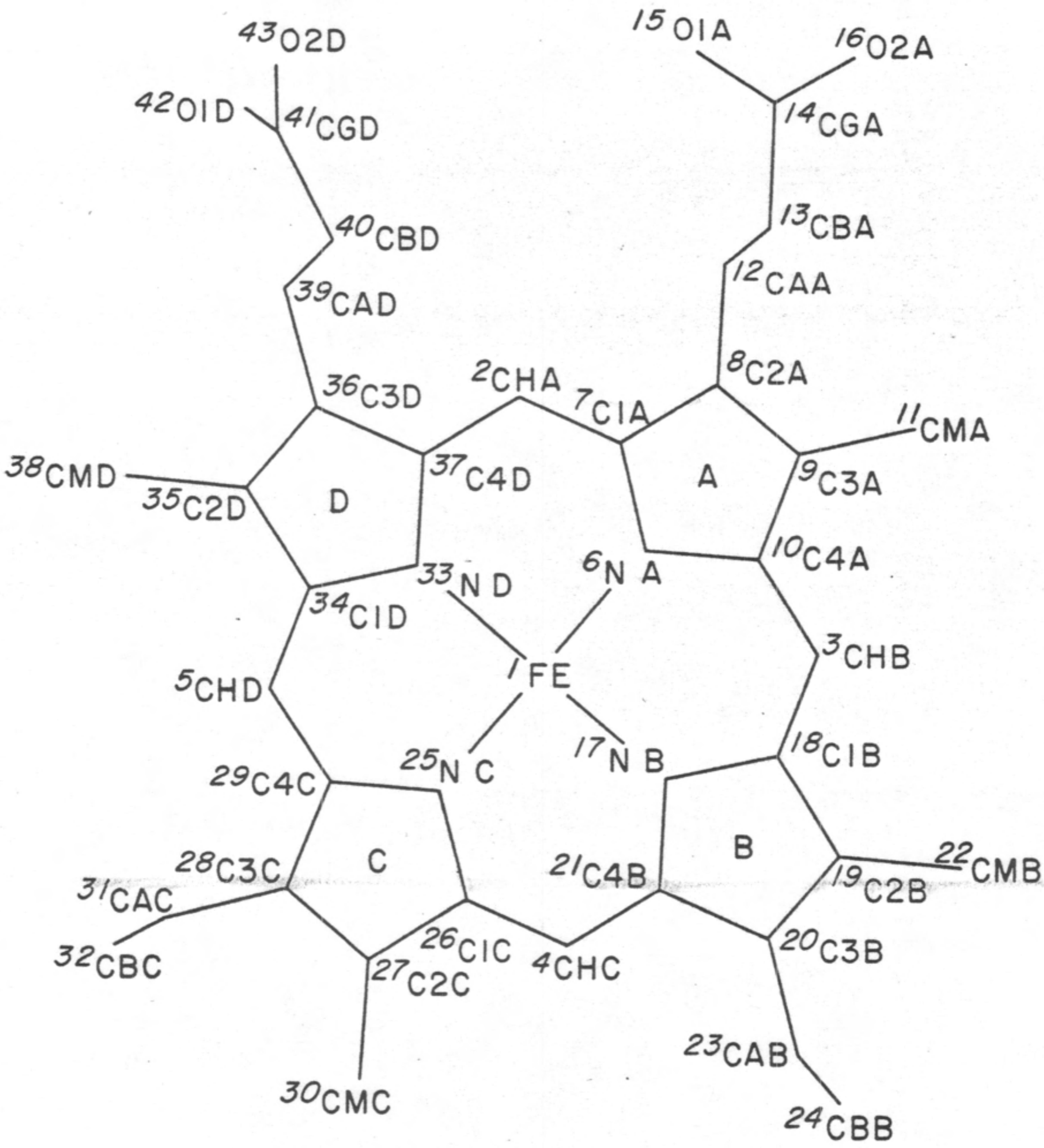


Uridine



C. <u>Non standard (HET) groups</u>

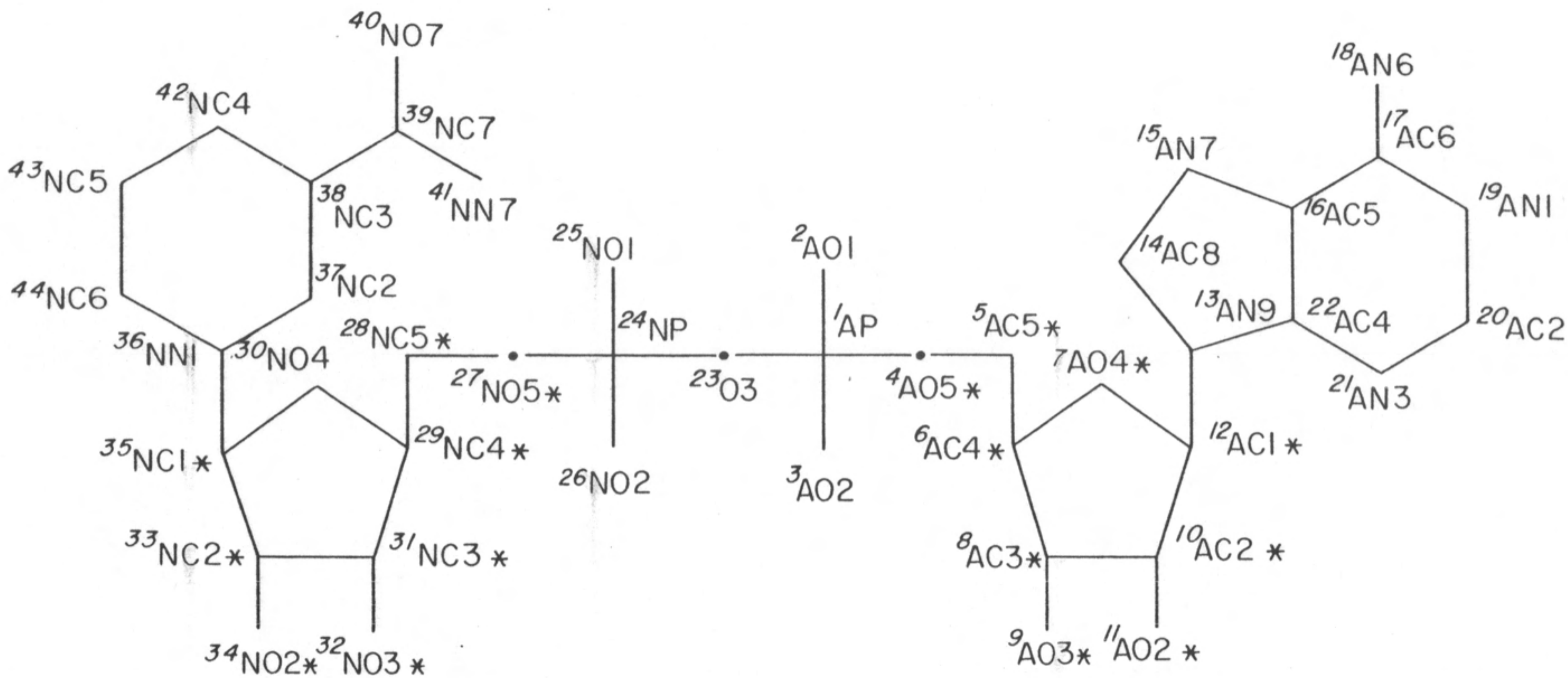
Because of the repeated occurrence of certain cofactors, prosthetic groups, etc., and the almost complete lack of uniformity in the nomenclature assigned by depositors, and in the absence of any authoritative precedent, the Data Bank has assigned a standard nomenclature and ordering of the atoms in some of these groups. These assignments appear on the following pages.



PROTEIN DATA BANK
STANDARD NOMENCLATURE FOR A HEME GROUP

(i) Order indicators are given as preceding superscripts.

(ii) Non-protein ligands of the iron atom are listed after those atoms given above.



Protein Data Bank Standard Nomenclature for an NAD Molecule

(i) Order indicators are given as preceding superscripts.

(ii) For an NADP molecule the atoms of the extra phosphate group will be listed after those above.

APPENDIX C

Residue Names and Abbreviations

A. Amino Acids

Residue	Abbr.	Syn.	Residue	Abbr.	Syn.
Acidic unknown	ACD		Homoserine	HSE	
Acetyl	ACE		Hydroxproline	HYP	
Alanine	ALA	A	Hydroxylysine	HYL	
β-Alanine	ALB		Isoleucine	ILE	I,ILU
Aliphatic unknown	ALI		Leucine	LEU	L
γ-Aminobutyric acid	ABU		Lysine	LYS	K
Arginine	ARG	R	Methionine	MET	M
Aromatic unknown	ARØ		Ornithine	ØRN	
Asparagine	ASN	N	Phenylalanine	PHE	F
Aspartic acid	ASP	D	Proline	PRØ	P,PR0,PRZ
ASP/ASN ambiguous	ASX	B	Pyrrolidone carboxylic acid	PCA	
Basic unknown	BAS		Sarcosine	SAR	
Betaine	BET		Serine	SER	S
Cysteine	CYS	C,CYH CSH	Taurine	TAU	
Cystine	CYS	C,CSS CYX	Threonine	THR	T
Formyl	FØR		Thyroxine	THY	
Glutamic acid	GLU	E	Tryptophan	TRP	W,TRY
Glutamine	GLN	Q	Tyrosine	TYR	Y
GLU/GLN ambiguous	GLX	Z	Unknown	UNK	X
Glycine	GLY	G	Valine	VAL	V
Histidine	HIS	H	Water	HØH	H2O,WAT,ØH2

Notes: 1) Standard residue abbreviations conform to the IUPAC-IUB rules in J. Biol. Chem., 241, 527, 2491 (1966).

2) Recognizable synonyms, such as those above, will be changed to the standard abbreviation.

3) Non-standard residues (metals, prosthetic groups, etc.) are given a three-character designation which is defined in a special HET record--see page 5.

4) To avoid confusion here within residue abbreviations, the alphabetic character is written "Ø" and the numeric "0". This convention is not observed elsewhere throughout these specifications.

B. Nucleic Acids

Abbreviations conform to the IUPAC-IUB recommendations (J. Biol. Chem., 251, 5171 (1970) for nucleosides with some extensions to cover the modified nucleosides and alterations because of character-set limitations.

Currently the following abbreviations are in use for the indicated residues.

Residue	Abbr
Adenosine	A
1-Methyladenosine	1MA
Cytidine	C
5-Methylcytidine	5MC
2'-O-Methylcytidine	ØMC
Guanosine	G
N(2)-Methylguanosine	2MG
N(2)-Dimethylguanosine	M2G
7-Methylguanosine	7MG
2'-O-Methylguanosine	ØMG
Ribosylthymine	5MU
Uridine	U
Dihydrouridine	H2U
Pseudouridine	PSU
Wybutosine	YG

Notes:

1) Non-standard residues (unusual modified bases, metals, etc.) are given a three-character designation which is defined in a special HET record--see page 5.

2) To avoid confusion here within residue abbreviations, the alphabetic character is written "Ø" and the numeric "0". This convention is not observed elsewhere throughout these specifications.

APPENDIX D

Protein Data Bank Conventions

In order to allow access to portions of the Data Bank entries over an interactive computer network, it has been decided to tighten the rules under which certain categories of information are presented. New specifications for the bibliographic citations given in the JRNL and REMARK 1 records are given in Appendix E. Concurrent with these changes it was deemed desirable to allow the availability of both upper and lower case characters on some computers to be exploited by inserting certain typesetting codes.

In addition to the detailed specifications given below the following general rules apply:

(i) No word is to be hyphenated and split over two records.

(ii) Only the surname of an author or editor is given in full, other names are indicated by initials only, e.g. A.B. Cooper.

(iii) Blanks and hyphens are used in author lists only if they are properly part of a name (e.g. C.-I.Branden, C.J.Birkett-Clews,L.Riva di Sansaverino).

(iv) The word Junior is written out in full.

(v) Author or Editor lists are terminated by a blank.

Typesetting codes are kept to a minimum by a judicious choice of default conventions. In the text strings of HEADER,COMPND, SOURCE, REF, TITL and PUBL records all letters are lower case unless preceded by one of the following characters-blank, comma, period, left parenthesis or asterisk. The occurrence of a slash forces all succeeding letters to be upper case until column 70 is reached or either a dollar sign or a hyphen (minus sign) is encountered.

Superscripts are initiated and terminated by double equals signs e.g. S==2+==.

Subscripts are initiated and terminated by single equals signs e.g. F=c=.

For author lists all characters are lower case unless they are adjacent to a period, or a comma, or preceded by an asterisk. A dollar sign is used to separate a lower-case character from a period or comma which otherwise would force upper case.

Comments for specific record types follow:

1. HEADER. In cols 11-50 of these records an attempt is made to assign the macromolecule to some functional class. No general classification scheme for biological macromolecules according to function yet exists (except for enzymes) and so the designation given here is intended to be informative rather than definitive. Its future use in indexing and subdividing the file is envisioned.

2. COMPND
3. SOURCE
4. AUTHOR

For these three records, the text portion of continuation lines begins in col. 12, leaving col. 11 blank. Such continuation lines are numbered 2,3 etc. in col. 10. The first line in each of these records has col. 10 blank.

5. JRNL. If the coordinate set held is identified in the literature, the paper containing the definition is cited here. If an article defines more than one coordinate set, the particular designation assigned is given in the REMARKs section. The format of literature citations for both the JRNL and REMARK 1 records is given in Appendix E.

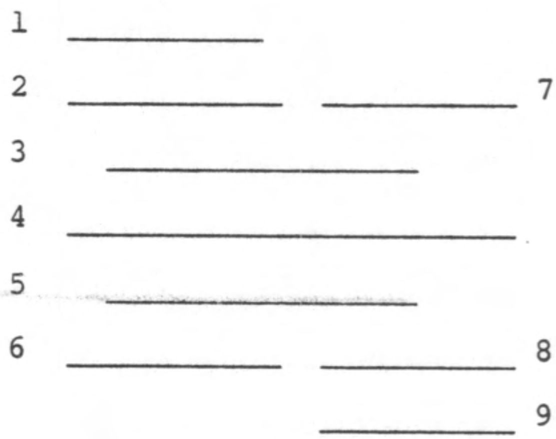
6. REMARK. The first REMARK lists the important papers relating to a structure which originate from the depositor's laboratory. These papers are usually listed in inverse chronological order, except if a particular article (or series of articles) is considered to be a definitive description, in which case it may appear first. If any citation is given in the JRNL records, it is not repeated here. References to the Atlas of Protein Sequence and Structure, and to the Atlas of Macromolecular Structure on Microfiche have been included where appropriate.

   REMARKS 2 and 3 are reserved for statements relating to the resolution and refinement of the structure analysis. Other general commentary is given in subsequent numbered REMARKs.

7. SEQRES. This set of records gives the number and sequence of residues in each chain of the particular macromolecule or complex under study. No cognizance of homologous molecules on which the residue sequence identifiers may be based is taken here. Residues which are present but not found in the crystallographic analysis are listed, but residues removed from the chain terminii (e.g., during an activation process) are not included. Residues excised from the chain (not at the terminii, e.g., in α-chymotrypsin) are represented by EXC in the SEQRES records.

   In general if the macromolecule is composed of two or more chains which are commonly conceptualised as being logically separable, e.g., Ribonuclease S, or Papain with an oligopeptide inhibitor, then separate sets of SEQRES records are provided for each of these chains. If however these chains are usually thought of as comprising an integral unit (e.g. the three chains of alpha-chymotrypsin) a single set of SEQRES records are given.

12. SHEET. For the case of bifurcated sheets, or those containing split strands (i.e., one strand comprised of two distinct amino acid runs), sufficient redundant sheets are defined to accommodate the bifurcations. For the case illustrated below two sheets would be given:



In this sketch each line represents a contiguous amino-acid run forming a strand of a sheet.

The strands labelled 1, 2, 3, 4, 5, 6 would comprise one sheet and those denoted 7, 3, 4, 5, 8, 9 another. This redundancy would be explicitly noted in a REMARK.

13. TURN. These records were originally set up to describe four-residue turns (β turns) but three residue turns (γ turns) may be included with a notation.

14. SSBOND. Each pair of cysteine residues which participate in a disulfide bond is listed here. Intra-chain bonds are listed before inter-chain linkages. The amino-acid with the lower sequence identifier is listed first in each intra-chain pair. For inter-chain pairs the cysteine which occurs first in the Data Bank entry is listed first.

15. CRYST1. The unit cell dimensions of the native crystals are given here unless explicitly stated otherwise. Native in this context means "underivatized" but if a derivative structure is solved as the native, e.g., tosyl elastase, then the cell dimensions of this pseudo-native macromolecule are given.

The Hermann-Mauguin space-group symbol is given without parentheses or slashes e.g., P 43 21 2.

Confusion over the value to use for Z (number of molecules per cell) arises because of different conceptions of the meaning of "molecule." We have adopted the (crystallographic) convention that Z should equal the number of times the same polymeric chain is contained in the cell. In the case of different numbers of different chains per cell this will be explained in the REMARK section and Z will denote the number of the more populous species per cell.

18. ATOM. The orthogonal Ångstrom coordinates stored are either those specified by the depositor or defined with respect to the default set of orthogonal axes (Appendix A). In the case that the stored coordinates are in orthogonal Ångstroms but not with respect to the default axial system then this is explained in a REMARK. The occupancy and temperature factor fields will contain the default values 1.0 and 0.0 if these parameters were not deposited. Otherwise these fields will contain the supplied quantities in their original form, i.e., as fractional occupancy/isotropic thermal-vibration parameter (B) or electron count/atomic-radius form. If an atom is found in two or more locations (i.e., disordered) the records carrying the different coordinates for the atom in question occur together.

19. HETATM. Comments as above for the ATOM records apply. In order to avoid problems associated with the special characters ' and ", which are often employed for saccharide atomic nomenclature, the more standard characters * and $ are employed instead. A standard nomenclature and ordering (this may not be the same as that employed by the depositor) for the atoms of all non-standard groups is assigned. This nomenclature is illustrated for some commonly-occuring non-standard groups in Appendix B.

20. TER. These records are inserted after the carboxy-terminal (3'- terminal) residue of each polypeptide (nucleotide) chain if the terminal residue is represented in the data set. TER cards are also inserted to denote the ends of inhibitors or pseudo-substrates, which are obtained by condensing like structural units present (e.g., peptides, oligonucleotides, oligosaccharides, etc.).

21. CONECT. These records may be used to specify all linkages not implied by the primary structure. Bonds from the polymeric chain to any non-standard groups present are given here as are all covalent bonds within such groups. Cross-links between polymeric chains (e.g., disulfide bonds) are specified as are any other important linkages deemed worthy of inclusion by the depositor. The connectivity list given here is redundant in that each bond indicated is given twice, once with each of the two atoms involved specified in cols 7-11. These CONECT records occur in increasing order of the atom serial numbers they carry in cols 7-11. The target-atom serial numbers carried on these records also occur in increasing order.

APPENDIX E

Formats for Literature Citations

References to published works from the depositor's laboratory and relating to the Data Bank Entry may be carried in either the JRNL or REMARK 1 records. The subsidiary tag-words AUTH, TITL, EDIT, PUBL, and REFN are used as appropriate to indicate the information carried. The details of these specifications are identical for the JRNL and REMARK 1 records except that for each citation in the latter list a lead record is provided which carries the word REFERENCE in cols 12-20 and a left-justified ordinal in cols 22-23. The details are exemplified by a JRNL citation.

Cols	1-4	JRNL
	13-16	AUTH (or EDIT)
	17-18	Continuation record number - blank for the first AUTH record of this citation - set to 2,3,etc., for succeeding records.
	20-70	Author list or Editor list.
Cols	1-4	JRNL
	13-16	TITL
	17-18	Continuation record number
	20-70	Title of Article
Cols	1-4	JRNL
	13-15	REF
	17-18	Continuation record number
	20-47	Name of publication (including section or series designation) (See Appendix F)
	50-51	V.
	53-55	Volume number
	57-61	First page number of article
	63-66	Year of publication

If more than one REF record is necessary to carry the name of the publication the volume number, page and date of publication is always carried on the first record.

Cols	1-4	JRNL
	13-16	PUBL (this category is omitted for journal articles)
	17-18	Continuation record number
	20-70	Name of publisher and city of publication
Cols	1-4	JRNL
	13-16	REFN
	20-23	ASTM
	25-30	Code from ASTM list (See Appendix F)
	33-34	Country of publication (journals only)
	36-39	ISSN or ISBN
	41-65	ISSN or ISBN number (See Appendix F)
	68-70	Code from Cambridge Crystallographic Data Centre list (See Appendix F)

APPENDIX F

Names and Identifying Codes for Selected Publications

JOURNAL	ASTM CODE	COUNTRY	ISSN CODE	CAMBRIDGE CODE
ACC.CHEM.RES.	ACHRE4	US	ISSN 0001-4842	411
ACTA CRYSTALLOGR.	ACCRA9	DK	ISSN 0001-5520	001
ACTA CRYSTALLOGR.,SECT.A	ACACBN	DK	ISSN 0001-5520	108
ACTA CRYSTALLOGR.,SECT.B	ACBCAR	DK	ISSN 0001-5520	107
ADV.EXP.MED.BIOL.	AEMBAP	US	ISSN 0065-2598	412
ADV.PROTEIN CHEM.	APCHA2	US	ISSN 0065-3233	433
ANN.REV.BIOCHEM.	ARBOAW	US	ISSN 0066-4154	413
BIOCHEM.	BICHAW	US	ISSN 0006-2960	033
BIOCHEM.BIOPHYS.RES.COMM.	BBRCA9	US	ISSN 0006-291X	146
BIOCHEM.J.	BIJOAK	UK	ISSN 0306-3275	043
BIOCHIM.BIOPHYS.ACTA	BBACAQ	NE	ISSN 0006-3002	113
BIOPHYS.STRUC.MECHANISM	BSMHBH	GW	ISSN 0340-1057	414
CAN.J.BIOCHEM.	CJBIAE	CN	ISSN 0008-4018	415
COLD SPRING HARBOR SYMP. QUANT.BIOL.	CSHSAZ	US	ISSN 0069-617X	421
DISS.ABSTR.B	DABBBA	US	ISSN 0419-4217	101
ESSAYS IN BIOCHEM.	ESBIAV	UK	ISSN 0071-1365	416
EUR.J.BIOCHEM.	EJBCAI	IX	ISSN 0014-2956	262
/FEBS$ LETT.	FEBLAL	NE	ISSN 0014-5793	165
J.AM.CHEM.SOC.	JACSAT	US	ISSN 0002-7863	004
J.BIOCHEM.	JOBIAO	JA	ISSN 0021-924X	418
J.BIOL.CHEM.	JBCHA3	US	ISSN 0021-9258	071
J.MOL.BIOL.	JMOBAK	UK	ISSN 0022-2836	070
NATURE	NATUAS	UK	ISSN 0028-0836	006
NATURE NEW BIOL.	NNBYA7	UK	ISSN 0028-0836	192
NATURWISS.	NATWAY	GW	ISSN 0028-1042	049
NUCLEIC ACIDS RESEARCH	NARHAD	UK	ISSN 0305-1048	389
PHIL.TRANS.R.SOC.LONDON, SER.A	PTRMAD	UK	ISSN 0080-4614	152
PHIL.TRANS.R.SOC.LONDON, SER.B	PTRBAE	UK	ISSN 0080-4622	441
PROC.NAT.ACAD.SCI.USA	PNASA6	US	ISSN 0027-8424	040
PROC.R.SOC.LONDON,SER.A	PRLAAZ	UK	ISSN 0080-4630	014
PROC.R.SOC.LONDON,SER.B	PRLBA4	UK	ISSN 0080-4649	338
PROG.STEREOCHEM.	PRSTAP	US	ISSN 0079-6808	419
SCIENCE	SCIEAS	US	ISSN 0036-8075	038
SCI.AM.	SCAMAC	US	ISSN 0036-8733	420
TO BE PUBLISHED				353

BOOK NAME		ISBN CODE	CAMBRIDGE CODE
ANN.N.Y.ACAD.SCI.	ANYAA9	ISBN SEE EACH BOOK	332
ATLAS OF MACROMOLECULAR STRUCTURE ON MICROFICHE		ISBN 0-917934-01-6	434
ATLAS OF PROTEIN SEQUENCE AND STRUCTURE (1972)		ISBN 0-912466-02-2	435
ATLAS OF PROTEIN SEQUENCE AND STRUCTURE,SUPPLEMENT 1		ISBN 0-912466-04-9	435
ATLAS OF PROTEIN SEQUENCE AND STRUCTURE,SUPPLEMENT 2		ISBN 0-912466-05-7	435
THE ENZYMES		ISBN SEE EACH BOOK	436

APPENDIX G

Formulas and Molecular Weights for Standard Residues

Residue Name	Code	Formula	Mol. wt.
Amino Acids			
Acetic acid	ACE	$C_2H_4O_2$	60.05
Alanine	ALA	$C_3H_7N_1O_2$	89.09
Arginine	ARG	$C_6H_{14}N_4O_2$	174.20
Asparagine	ASN	$C_4H_8N_2O_3$	132.12
Aspartic acid	ASP	$C_4H_7N_1O_4$	133.10
Cysteine	CYS	$C_3H_7N_1O_2S_1$	121.15
Formic acid	FØR	$C_1H_2O_2$	46.03
Glutamine	GLN	$C_5H_{10}N_2O_3$	146.15
Glutamic acid	GLU	$C_5H_9N_1O_4$	147.13
Glycine	GLY	$C_2H_5N_1O_2$	75.07
Histidine	HIS	$C_6H_9N_3O_2$	155.16
Isoleucine	ILE	$C_6H_{13}N_1O_2$	131.17
Leucine	LEU	$C_6H_{13}N_1O_2$	131.17
Lysine	LYS	$C_6H_{14}N_2O_2$	146.19
Methionine	MET	$C_5H_{11}N_1O_2S_1$	149.21
Pyrollidone carboxylic acid	PCA	$C_5H_7N_1O_3$	129.12
Phenylalanine	PHE	$C_9H_{11}N_1O_2$	165.19
Proline	PRØ	$C_5H_9N_1O_2$	115.13
Serine	SER	$C_3H_7N_1O_3$	105.09

Residue Name	Code	Formula	Mol. wt.
Threonine	THR	$C_4H_9N_1O_3$	119.12
Tryptophan	TRP	$C_{11}H_{12}N_2O_2$	204.23
Tyrosine	TYR	$C_9H_{11}N_1O_3$	181.19
Valine	VAL	$C_5H_{11}N_1O_2$	117.15
ASP or ASN	ASX		132.61
GLU or GLN	GLX		146.64
Undetermined	UNK		128.16
Water	HØH	H_2O_1	18.015

Residue Name	Code	Formula	Mol. wt.
Nucleotides (RNA)			
Adenosine	A	$C_{10}H_{14}N_5O_7P_1$	347.22
Cytidine	C	$C_9H_{14}N_3O_8P_1$	323.20
Guanosine	G	$C_{10}H_{14}N_5O_8P_1$	363.22
Uridine	U	$C_9H_{13}N_2O_9P_1$	324.18
1-Methyladenosine	1MA	$C_{11}H_{16}N_5O_7P_1$	361.25
5-Methylcytidine	5MC	$C_{10}H_{16}N_3O_8P_1$	337.23
2'-O-Methylcytidine	ØMC	$C_{10}H_{17}N_3O_8P_1$	338.23
N(2)-Methylguanosine	2MG	$C_{11}H_{10}N_5O_8P_1$	377.25
N(2)-Dimethylguanosine	M2G	$C_{12}H_{18}N_5O_8P_1$	391.28
7-Methylguanosine	7MG	$C_{11}H_{16}N_5O_8P_1$	377.25
2'-O-Methylguanosine	ØMG	$C_{11}H_{16}N_5O_8P_1$	377.25
Ribosylthymidine	5MU	$C_{10}H_{16}N_2O_{10}P_1$	355.22
Dihydrouridine	H2U	$C_9H_{15}N_2O_9P_1$	326.20
Pseudouridine	PSU	$C_9H_{13}N_2O_9P_1$	324.18
Wybutosine	YG	$C_{21}H_{28}N_6O_{12}P_1$	587.48