Protein Data Bank Atomic Coordinate and Bibliographic Entry Format Description

Macromolecular Atomic Coordinate and Bibliographic File Entry

For each data entry, the file consists of records each of 80 characters. The record sequence is as follows:

HEADER	:	Date entered into Data Bank; identification code
OBSLTE	:	Identifies entries which have been replaced
COMPND	:	Name of molecule and identifying information
SOURCE	:	Species, organ, tissue, and mutant from which the molecule has been obtained, where applicable
AUTHOR	:	Names of contributors
REVDAT	:	Revision date; identifies current modification level
SPRSDE	:	Identifies entries which have replaced others
JRNL	:	Literature citation that defines coordinate set
REMARK	:	General remarks
SEQRES	:	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	:	Identification 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
TVECT	:	Translation vector for infinite covalently connected structures
ATOM	:	Atomic coordinate records for "standard" groups
HETATM	:	Atomic coordinate records for "non-standard" groups
SIGATM	:	Standard deviations of atomic parameters
ANISOU	:	Anisotropic temperature factors
SIGUIJ	:	Standard deviations of anisotropic temperature factors
TER	:	Chain terminator
CONECT	:	Connectivity records
MASTER	:	Master control record with checksums of total number of records in the file, for selected record types
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.

Records are present in each entry in the order specified above with the following exceptions:

(i) ATOM and HETATM records appear in the order appropriate to the structure.

(ii) TER records may appear among ATOM or HETATM records as appropriate.

(iii) SIGATM, ANISOU and SIGUIJ records, when present, directly follow the corresponding ATOM (or HETATM) record in the order SIGATM, ANISOU, SIGUIJ.

Note: Over the course of the next few years we anticipate making some changes in the format of our atomic coordinate entries. As far as possible, we will do this by adding new types of records; user programs should thus be prepared to handle or ignore record types that are not currently defined. Any revisions to existing record types will be announced well in advance via the Newsletter. Because of user requests, we are reserving a class of records for users' definition and use. All records beginning with the four letters USER are reserved for user definition and will be ignored by our programs.

RECORD FORMATS

1. **HEADER**	Cols.	1-6	HEADER
		11-50	Functional classification of macromolecule
		51-59	Date of deposition into Data Bank[(i)]
		63-66	Identification code[(ii)]

FORMAT (6A1,4X,40A1,9A1,3X,A4)

Note: (i) The date is given in the form dd-mmm-yy (e.g., December 1, 1983 is given as 01-DEC-83).

(ii) Each macromolecule is assigned an identification code. The code consists of 1 numeric and 3 alphanumeric characters.

2. **OBSLTE**	Cols.	1-6	OBSLTE
		9-10	Continuation field (this field will be blank for the first OBSLTE record in each entry and numbered 2, 3, etc. for continuation records)
		12-20	Date this entry was replaced
		22-25	Identification code of this entry which is now obsolete
		32-35	Identification code of a new entry which has replaced this old entry
		37-40	
		⋮	
		67-70	Identification code of a new entry which has replaced this old entry

FORMAT (6A1,2X,2A1,1X,9A1,1X,4A1,5X,8(1X,4Al))

Note: This record will be inserted only in archived entries that are no longer distributed.

3. **COMPND**	Cols.	1-6	COMPND
		11-70	Name of macromolecule

FORMAT (6A1,4X,60A1)

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. The Enzyme Commission numbers are obtained from Enzyme Nomenclature 1978, Published for the International Union of Biochemistry by Academic Press, Inc.

4. **SOURCE** Cols.

1-6	SOURCE
11-70	Species, organ, tissue, and mutant from which the macromolecule has been obtained. The systematic name of the species is given in parentheses.

FORMAT (6A1,4X,60A1)

5. **AUTHOR** Cols.

1-6	AUTHOR
11-70	Name(s) of contributor(s)

FORMAT (6A1,4X,60A1)

6. **REVDAT** Cols.

1-6	REVDAT
8-10	Modification number$^{(i)}$
11-12	Continuation field$^{(ii)}$
14-22	Date$^{(iii)}$
24-28	Identification name used for the correction
32	Modification type$^{(iv)}$
40-70	Record types that were corrected.

FORMAT (6Al,1X,I3,2A1,1X,9A1,1X,5A1,3X,I1,7X,31Al)

Notes:

(i) Each revision will be given a modification number assigned in increasing numerical order but inserted in the entry in decreasing numerical order. New entries will be assigned the modification number 1.

(ii) For each modification, more than one REVDAT record may be supplied. This field will be blank on the first record of each modification, and numbered 2, 3, etc. for continuation records.

(iii) For new entries this date will be the date when the entry was released for distribution rather than the date of deposition which appears in the HEADER record.

(iv) The following integer values will be used to identify the modification type: (In case of revisions with more than one possible type, the highest value applicable will be assigned.)

3 – Used for modifications affecting the coordinates or their transforms. To be used for entries with revisions to any of the following records.

a CRYST1

b ORIGX

c SCALE

d MTRIX

e TVECT

f ATOM item HETATM

g SIGATM

2 – Used for modifications to the CONECT records.

1 – Used for all other types of modifications, mainly typographical in nature.

0 – Initial entry. New entries will contain a REVDAT record flagged as modification type 0.

7. **SPRSDE** Cols.

1-6	SPRSDE
9-10	Continuation field (this field will be blank for the first SPRSDE record in each entry and numbered 2, 3, etc. for continuation records).
12-20	Date that this entry superseded an older one
22-25	Identification code of this entry which is replacing an older one
32-35	Identification codes of the entries which are being replaced
67-70	Identification codes of the entries which are being replaced

FORMAT (6A1,2X,2A1,1X,9A1,1X,4A1,5X,8(1X,4A1))

8. **JRNL** Cols.

1-4	JRNL
11-70	Literature citation that defines the coordinate set

FORMAT (6A1,4X,60A1)

Note: See appendices D and E for detailed specifications.

9. **REMARK** Cols.

1-6	REMARK
8-10	Remark number
12-70	Text of remark

FORMAT (6A1,1X,I3,59A1)

Note: The first REMARK has serial number 1, the second 2, etc. See appendices D and E for detailed specifications.

10. **SEQRES** Cols.

1-6	SEQRES
9-10	Serial number of SEQRES record for current chain
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))

11. **FTNOTE** Cols.

Cols.	
1-6	FTNOTE
8-10	Footnote number
12-70	Footnote text

FORMAT (6A1,1X,I3,1X,59A1)

Note: 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.

12. **HET** Cols.

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. The Protein Data Bank attempts to use uniform atom nomenclature for HET groups, as illustrated for commonly occurring groups in Appendix B. All non-standard groups (i.e., those not assigned a standard code in Appendix C) are defined in HET records. If there is insufficient space in the text portion of the HET record to properly define a non-standard group then the definition will be given in a REMARK and referenced here. An entry of -999 in the sequence number field is used to indicate that the HET group occurs too frequently, in the present entry, to use a separate HET record for each occurrence.

13. **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 may be 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 04 FE1 +++).

14. **HELIX** Cols. 1-6 HELIX

8-10 Serial number (helix number)

12-14 Helix identifier (rightjustified)(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 residue.

(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	10	Polyproline

15. **SHEET** Cols.

1-5 SHEET

8-10 Strand number(i)(v)

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

15-16 Number of strands

Columns	Field	Group
18-20	Residue name	Initial residue(ii)
22	Chain identifier	
23-26	Residue seq. no.	
27	Insertion code	
29-31	Residue name	Terminal residue
33	Chain identifier	
34-37	Residue seq. no.	
38	Insertion code	
39-40	Sense of this strand with respect to previous strand(iii)	
42-45	Atom name	In current strand. Registration(iv)
46-48	Residue name	
50	Chain identifier	
51-54	Residue seq. no.	
55	Insertion code	
57-60	Atom name	In previous strand. Registration(iv)
61-63	Residue name	
65	Chain identifier	
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.

16. **TURN** Cols.

Cols.	Content	
1-5	TURN	
8-10	Sequence number (Turn number)	
12-14	Turn identifier (3 characters)	
16-18	Residue name	Residue i
20	Chain identifier	Residue i
21-24	Residue seq. no.	Residue i
25	Insertion code	Residue i
27-29	Residue name	Residue i+3 (or i+2 for γ bend)
31	Chain identifier	Residue i+3 (or i+2 for γ bend)
32-35	Residue seq. no.	Residue i+3 (or i+2 for γ bend)
36	Insertion code	Residue i+3 (or i+2 for γ bend)
41-70	Comment	

FORMAT (6A1,1X,I3,1X,A3,1X,A3,1X,A1,I4,A1,1X,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.

17. <u>**SSBOND**</u> 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 seq. no.
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)

18. <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	
19-21	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, these may be specified on additional records bearing the same site identifier.

19. **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)

20. **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 submitted coordinated be X_{sub}, Y_{sub}, Z_{sub} and the orthogonal Å coordinates contained in the file be X,Y,Z. Then

$$X_{sub} = O_{11}X + O_{12}Y + O_{13}Z + T_l$$
$$Y_{sub} = O_{21}X + O_{22}Y + O_{23}Z + T_2$$
$$Z_{sub} = O_{31}X + O_{32}Y + O_{33}Z + T_3$$

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.

21. **SCALE**

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 $\underline{x}_{frac}$, $\underline{y}_{frac}$, $\underline{z}_{frac}$. Then

$$\begin{aligned}\underline{x}_{frac} &= S_{11}X + S_{12}Y + S_{13}Z + U_1\\ \underline{y}_{frac} &= S_{21}X + S_{22}Y + S_{23}Z + U_2\\ \underline{z}_{frac} &= S_{31}X + S_{32}Y + S_{33}Z + U_3\end{aligned}$$

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

The standard orthogonal Å coordinate system is related to the axial system of the unit cell supplied (CRYST1 record) by the definition below. (Non-standard coordinate systems are generally explained in the REMARKs.)

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 $\vec{C}$ x $\vec{A}$

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

Appendix A details the derivation of this coordinate transformation.

22. **MTRIX**

Cols. 1-6	8-10	11-20	21-30	31-40	46-55	59-60
MTRIX1	Ser.no.	M_{11}	M_{12}	M_{13}	V_1	IGIVEN
MTRIX2	Ser.no.	M_{21}	M_{22}	M_{23}	V_2	IGIVEN
MTRIX3	Ser.no.	M_{31}	M_{32}	M_{33}	V_3	IGIVEN

FORMAT (6A1,1X,I3,3F10.5,5X,F10.5,3X,I2)

Notes: (i) One trio of MTRIX records with a constant serial number is given for each non-crystallographic symmetry operation that is defined.

(ii) The MTRIX transformations operate on the stored coordinates to yield equivalent representations of the molecule in the same space.

(iii) If coordinates for the representations which are approximately related by the transformation in question are contained in the file, the quantity IGIVEN is set to 1. Otherwise this indicator carries zero (blank).

23. **TVECT** Cols.

1-5	TVECT
8-10	Serial number
11-20 21-30 31-40	component of translation vector
41-70	Comment

FORMAT (6A1,1X,I3,3F10.5,30A1)

Note: For structures not comprised of discrete molecules (e.g., infinite polysaccharide chains) the Protein Data Bank entry will contain a fragment which can be built into the full structure by the simple translation vectors of TVECT records.

24. **ATOM** Atomic coordinate records for "standard" groups

HETATM Atomic coordinate records for "non-standard" groups

Cols.	1-4	ATOM
or	1-6	HETATM
	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 seq. no.
	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(vi)
	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) See appendix B

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

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

(v) HETATM records are used for water molecules and atoms contained in HET groups.

(vi) If anisotropic temperature factors have been provided, the temperature factor field of the corresponding ATOM or HETATM record will contain the equivalent U-isotropic (U(eq)) which is calculated by

$$\begin{aligned} \mathrm{U(eq)} &= 1/3[\mathrm{U}(1,1) + \mathrm{U}(2,2) + \mathrm{U}(3,3) + \\ &\quad \mathrm{U}(1,2)\mathrm{a}*\mathrm{b}*\mathrm{abcos}\gamma + \mathrm{U}(1,3)\mathrm{a}*\mathrm{c}*\mathrm{accos}\beta + \\ &\quad \mathrm{U}(2,3)\mathrm{b}*\mathrm{c}*\mathrm{bccos}\alpha] \end{aligned}$$

where (a,b,c) are the direct cell dimensions given in the CRYST1 record and (a*,b*,c*) are the reciprocal cell dimensions.

25. **SIGATM**	Cols.	1-6	SIGATM
		7-27	Identical to corresponding ATOM record
		31-38	
		39-46	Standard deviations of the stored coordinates (Å)
		47-54	
		55-60	Standard deviation of occupancy
		61-66	Standard deviation of temperature factor
		68-70	Footnote number

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

26. **ANISOU** Cols.

1-6	ANISOU
7-27	Identical to corresponding ATOM or HETATM record
29-35	U(1,1)
36-42	U(2,2)
43-49	U(3,3) Anisotropic temperature factors x 10^4 (Å^2)
50-56	U(1,2)
57-63	U(1,3)
64-70	U(2,3)

FORMAT (6A1,I5,1X,A4,A1,A3,1X,A1,I4,A1,1X,6I7)

Note: (i) If anisotropic temperature factors have been provided, the temperature factor field of the corresponding ATOM or HETATM record will contain the equivalent U-isotropic (U(eq)) which is calculated by

$$\begin{aligned} \mathrm{U(eq)} &= 1/3[\mathrm{U}(1,1) + \mathrm{U}(2,2) + \mathrm{U}(3,3) + \\ &\quad \mathrm{U}(1,2)\mathrm{a}*\mathrm{b}*\mathrm{abcos}\gamma + \mathrm{U}(1,3)\mathrm{a}*\mathrm{c}*\mathrm{accos}\beta + \\ &\quad \mathrm{U}(2,3)\mathrm{b}*\mathrm{c}*\mathrm{bccos}\alpha] \end{aligned}$$

where (a,b,c) are the direct cell dimensions given in the CRYST1 record and (a*,b*,c*) are the reciprocal cell dimensions.

(ii) The anisotropic temperature factors will be stored in the same coordinate frame as the atomic coordinate records.

27. **SIGUIJ** Cols.

1-6	SIGUIJ	
7-27	Identical to corresponding ATOM (or HETATM) record	
29-35	Sigma U(1,1)	
36-42	Sigma U(2,2)	Standard deviations of anisotropic temperature factors x 10^4 (Å^2)
43-49	Sigma U(3,3)	
50-56	Sigma U(1,2)	
57-63	Sigma U(1,3)	
64-70	Sigma U(2,3)	

FORMAT (6A1,I5,1X,A4,A1,A3,1X,A1,I4,A1,1X,6I7)

28. **TER** Cols.

1-3	TER
7-11	Serial number
18-20	Residue name
22	Chain identifier
23-26	Residue seq. no.
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.

29. **CONECT** Connectivity records

Cols.

1-6	CONECT	
7-11	Serial number	
12-16		
17-21		Covalent bond connectivity (serial numbers of bonded atoms)
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 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.

The occurrence of a negative atom serial number on a CONECT record denotes that a translationally equivalent copy (see TVECT records) of the target atom specified is linked to the origin atom of the record.

30. <u>**MASTER**</u> Col.

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
35-40	Number of TURN records
41-45	Number of SITE records
46-50	Number of coordinate transformation records (ORIGX+SCALE+MATRIX)
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, for selected record types.

31. <u>**END**</u> End-of-entry record

Cols. 1-3 END

FORMAT (6A1)

APPENDIX A – COORDINATE SYSTEMS AND TRANSFORMATIONS

The coordinates stored in the Data Bank give the atomic positions measured in Ångstroms along three orthogonal directions. Unless otherwise specified, the default axial system (detailed below) will be assumed.

If $\vec{a}$, $\vec{b}$, and $\vec{c}$ describe the crystallographic cell edges and $\vec{A}$, $\vec{B}$, and $\vec{C}$ are unit vectors in the default orthogonal Å system, then

a. $\vec{A}$, $\vec{B}$, $\vec{C}$ and $\vec{a}$, $\vec{b}$, $\vec{c}$ have the same origin.

b. $\vec{A}$ is parallel to $\vec{a}$.

c. $\vec{B}$ is parallel to $\vec{C}$ x $\vec{A}$.

d. $\vec{C}$ is parallel to $\vec{a}$ x $\vec{b}$ (i.e., c$*$).

The matrix which premultiplies the column vector of fractional crystallographic coordinates (x_{frac}, y_{frac}, z_{frac}) to yield coordinates in the $\vec{A}$, $\vec{B}$, $\vec{C}$ system, i.e., (X,Y,Z) is

$$\begin{bmatrix} a & b\cos\gamma & c\cos\beta \\ o & b\sin\gamma & c(\cos\alpha\text{-}\cos\beta\,\cos\gamma)/\sin\gamma \\ o & 0 & V/(ab\,\sin\gamma) \end{bmatrix}$$

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

If the submitted coordinates are either fractions of the unit cell edge or are with respect to the default orthogonal system, the ORIGX and SCALE transformations will be given default values.

In general the depositor will have supplied:

(i) The original submitted coordinates, i.e., $\vec{X}_{sub}$

(ii) A transformation from $\vec{X}_{sub}$ to the orthogonal Å coordinates stored in the Data Bank ($\vec{X}$), i.e., $O_{sub}\vec{X}_{sub} + \vec{T}_{sub} = \vec{X}$.

(iii) A transformation from $\vec{X}_{sub}$ to fractional crystallographic coordinates $\vec{x}_{frac}$

i.e., $S_{sub}\vec{X}_{frac} + \vec{U}_{sub} = \vec{x}_{frac}$

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

i.e., $M_{sub}\vec{X}_{sub} + \vec{V}_{sub} = \vec{X}_{sub}$

Note: The notation $\vec{X}_{sub}$ is used for the column vector X_{sub}, Y_{sub}, Z_{sub}, etc.

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 ($\vec{X}$)

$$\vec{X} = O_{sub}\vec{X}_{sub} + \vec{T}_{sub}$$

(ii) The ORIGX transformation from stored to original coordinates (O, $\vec{T}$).

$$\begin{array}{lrl}
\text{From above} & \vec{X} = & O_{sub}\vec{X}_{sub} + \vec{T}_{sub} \\
\text{whence} & O_{sub}\vec{X}_{sub} = & \vec{X} - \vec{T}_{sub} \\
\therefore & \vec{X}_{sub} = & O_{sub}^{-1}\vec{X} + (-O_{sub}^{-1}\vec{T}_{sub}) \\
\text{Thus} & O = & O_{sub}^{-1} \\
\text{and} & \vec{T} = & -O_{sub}^{-1}\vec{T}_{sub}
\end{array}$$

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

$$\begin{array}{lrl}
\text{From above} & \vec{x}_{frac} = & S_{sub}\vec{X}_{sub} + \vec{U}_{sub} \\
\text{but} & \vec{X}_{sub} = & O_{sub}^{-1}\vec{X} + (-O_{sub}^{-1}\vec{T}_{sub}) \\
\therefore & \vec{x}_{frac} = & S_{sub}[O_{sub}^{-1}\vec{X} + (-O_{sub}^{-1}\vec{T}_{sub})] + \vec{U}_{sub} \\
\text{i.e.,} & \vec{x}_{frac} = & S_{sub}O_{sub}^{-1}\vec{X} + (-S_{sub}O_{sub}^{-1}\vec{T}_{sub}) + \vec{U}_{sub} \\
\therefore & S = & S_{sub}O_{sub}^{-1} \\
\text{and} & \vec{U} = & -(S_{sub}O_{sub}^{-1}\vec{T}_{sub}) + \vec{U}_{sub}
\end{array}$$

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

$$
\begin{aligned}
& & \vec{X}'_{sub} &= M_{sub}\vec{X}_{sub} + \vec{V}_{sub} \\
& & \vec{X}' &= O_{sub}\vec{X}'_{sub} + \vec{T}_{sub} \\
& & &= O_{sub}\{M_{sub}\vec{X}_{sub} + \vec{V}_{sub}\} + \vec{T}_{sub} \\
&\text{but} & \vec{X}_{sub} &= O_{sub}^{-1}\vec{X} + (-O_{sub}^{-1}\vec{T}_{sub}) \\
&\therefore & \vec{X}' &= O_{sub}\{M_{sub}[O_{sub}^{-1}\vec{X} + (-O_{sub}^{-1}\vec{T}_{sub})] + \vec{V}_{sub}\} + \vec{T}_{sub} \\
&\therefore & M &= O_{sub}M_{sub}O_{sub}^{-1} \\
&\text{and} & \vec{V} &= -O_{sub}M_{sub}O_{sub}^{-1}\vec{T}_{sub} + O_{sub}\vec{V}_{sub} + \vec{T}_{sub}
\end{aligned}
$$

In summary the stored coordinates and transformations are:

$\vec{X}$	(ATOM, HETATM records)
O, $\vec{T}$	(ORIGX records)
S, $\vec{U}$	(SCALE records)
M, $\vec{V}$	(MTRIX records)

APPENDIX B - ATOM NAMES

A. Amino Acids

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

(i) Greek letter remoteness codes are transliterated as follows:
α-A, β-B, γ-G, δ-D, ε-E, ζ-Z, η-H

(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)

(iii) For protein coordinate sets containing hydrogen atoms, the IUPAC-IUB rules[1] have been followed. Exceptions to these rules may occur in certain data sets at the depositors' request. Any such exceptions will be delineated clearly in FTNOTE and REMARK records.

---

[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.

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 avoided 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 this would result in identical atom names. In this case an alphabetic character is used to avoid ambiguity.

(iv) The ring-oxygen atom of the ribose is denoted 04 rather than 01 .

(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.

C. Non-Standard (HET) Groups

Because of the repeated occurrence of certain cofactors, prosthetic groups, etc., 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 subsequent pages, for the following groups:

APPENDIX C – RESIDUE NAMES AND ABBREVIATIONS

A. Amino Acids

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

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 pp. 6-7.

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. 245, 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
l-Methyladenosine	1MA
Cytidine	C
5-Methylcytidine	5MC
2'-0-Methylcytidine	ØMC
Guanosine	G
l-Methylguanosine	lMG
N(2)-Methylguanosine	2MG
N(2)-Dimethylguanosine	M2G
7-Methylguanosine	7MG
2'-0-Methylguanosine	ØMG
Wybutosine	YG
Thymidine	T
Uridine	U
Modified Uridine	+U
Dihydrouridine	H2U
Ribosylthymidine	5MU
Pseudouridine	PSU

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

C. Miscellaneous

The following residue names are used to identify other commonly occurring groups.

Residue	Abbr.	Synonym
Acetyl	ACE	
Formyl	FØR	
Water	HØH	H2Ø,WAT,ØH2
Unknown	UNK	

Note: To avoid confusion here within residue abbreviations, the alphabetic character is written "Ø" and the numeric "O". 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. 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 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 equal signs, e.g.,S == 32+ ==.

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 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 higher 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 conceptualized 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 α-chymotrypsin) a single set of SEQRES records is given.

   If the residue sequence is unknown, the number of residues thought to comprise the molecule is entered in cols. 14-17, col. 10 contains '0' and cols. 20-22 contain 'UNK'.

8. **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:

```
1  ----------
2  -----------  ---------- 7
3    ----------------
4  ------------------------
5    ----------------
6  -----------  ---------- 8
                ---------- 9
```

   In this sketch each line represents a continuous amino-acid run forming 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.

9. **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.

10. **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.

11. **CRYSTl.** 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.

12. **TVECT.** These records are used to denote the translation vectors which are used to build the infinite covalently connected structure of which the Protein Data entry is representative.

13. **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.

14. **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 in their place. A uniform 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-occurring non-standard groups in Appendix B.

15. **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.).

16. 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)*
	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*
	33-34	Country of publication
	36-39	ISSN or ISBN
	41-65	ISSN or ISBN number*
	68-70	Code from Cambridge Crystallography Data Centre CODEN list*

*Note: A complete list of journal names and codes assigned to them is available upon request.

APPENDIX F – FORMULAS AND MOLECULAR WEIGHTS FOR STANDARD AMINO ACIDS AND NUCLEOTIDES

Note that these weights and formulas correspond to the unpolymerized state of the component. The elements of one water molecule are eliminated for each two components joined.

Name	Code	Formula	Mol. wt.
Amino Acids			
Alanine	ALA	C3 H7 N1 O2	89.09
Arginine	ARG	C6 H14 N4 O2	174.20
Asparagine	ASN	C4 H8 N2 O3	132.12
Aspartic acid	ASP	C4 H7 N1 O4	133.10
ASP/ASN ambiguous	ASX	C4 H7$\frac{1}{2}$ N1$\frac{1}{2}$ O3$\frac{1}{2}$	132.61
Cysteine	CYS	C3 H7 N1 O2 S1	121.15
Glutamine	GLN	C5 H10 N2 O3	146.15
Glutamic acid	GLU	C5 H9 N1 O4	147.13
GLU/GLN ambiguous	GLX	C5 H9$\frac{1}{2}$ N1$\frac{1}{2}$ O3$\frac{1}{2}$	146.64
Glycine	GLY	C2 H5 N1 O2	75.07
Histidine	HIS	C6 H9 N3 O2	155.16
Isoleucine	ILE	C6 H13 N1 O2	131.17
Leucine	LEU	C6 H13 N1 O2	131.17
Lysine	LYS	C6 H14 N2 O2	146.19
Methionine	MET	C5 H11 N1 O2 S1	149.21
Phenylalanine	PHE	C9 H11 N1 O2	165.19
Proline	PRØ	C5 H9 N1 O2	115.13
Serine	SER	C3 H7 N1 O3	105.09
Threonine	THR	C4 H9 N1 O3	119.12
Tryptophan	TRP	C11 H12 N2 O2	204.23
Tyrosine	TYR	C9 H11 N1 O3	181.19
Valine	VAL	C5 H11 N1 O2	117.15
Undetermined	UNK	C5 H6 N1 O3 (approx.)	128.16

Name	Code	Formula	Mol. wt.
Nucleotides			
Adenosine	A	C10 H14 N5 O7 P1	347.22
1-Methyladenosine	1MA	C11 H16 N5 O7 P1	361.25
Cytidine	C	C9 H14 N3 O8 P1	323.20
5-Methylcytidine	5MC	C10 H16 N3 O8 P1	337.23
2'-0-Methylcytidine	∅MC	C10 H17 N3 O8 P1	338.23
Guanosine	G	C10 H14 N5 O8 P1	363.22
1-Methylguanosine	1MG	C11 H16 N5 O8 P1	377.25
N(2)-Methylguanosine	2MG	C11 H16 N5 O8 P1	377.25
N(2)-Dimethylguanosine	M2G	C12 H18 N5 O8 P1	391.28
7-Methylguanosine	7MG	C11 H10 N5 O8 P1	377.25
2'-0-Methylguanosine	∅MG	C11 H16 N5 O8 P1	377.25
Wybutosine	YG	C21 H26 N6 O11 P1	587.48
Inosine	I	C10 H13 N4 O8 P1	348.21
Thymidine	T	C10 H15 N2 O8 P1	322.21
Uridine	U	C9 H13 N2 O9 P1	324.18
Dihydrouridine	H2U	C9 H15 N2 O9 P1	326.20
Ribosylthymidine	5MU	C10 H16 N2 O10 P1	355.22
Pseudouridine	PSU	C9 H13 N2 O9 P1	324.18
Miscellaneous			
Acetic Acid	ACE	C2 H4 O2	60.05
Formic Acid	FOR	C1 H2 O2	46.03
Water	HOH	H2 O1	18.015