Protein Data Bank

Atomic Coordinate and Bibliographic Entry Format Description

January 1985
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Summary of Record Types and Their Sequence</td>
<td>1</td>
</tr>
<tr>
<td>Specifications of Contents of Records</td>
<td>3</td>
</tr>
<tr>
<td>Appendix A Coordinate Systems and Transformations</td>
<td>20</td>
</tr>
<tr>
<td>Appendix B Atom Names</td>
<td>23</td>
</tr>
<tr>
<td>Appendix C Residue Names and Abbreviations</td>
<td>33</td>
</tr>
<tr>
<td>Appendix D Protein Data Bank Conventions</td>
<td>36</td>
</tr>
<tr>
<td>Appendix E Formats for Literature Citations</td>
<td>40</td>
</tr>
<tr>
<td>Appendix F Formulas and Molecular Weights for Standard Residues</td>
<td>42</td>
</tr>
</tbody>
</table>
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

JRNLM : 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

HELI : 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
MATRIX : 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-yyyy
   (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. **OBSLTER** Cols. 1-6 OBSLTER

   9-10 Continuation field (this field will be blank
   for the first OBSLTER 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,4A1))

   **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 Names(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** (6A1,1X,I3,2A1,1X,9A1,1X,5A1,3X,I1,7X,31A1)

**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) CRST1
b) ORIGX
c) SCALE
d) MTRIX
e) TVECT
f) ATOM
g) HETATM
h) 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**

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Date that this entry superseded an older one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

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

8. **JRNL**

<table>
<thead>
<tr>
<th>Cols.</th>
<th>1-4</th>
<th><strong>JRNL</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>11-70</td>
<td>Literature citation that defines the coordinate set</td>
<td></td>
</tr>
</tbody>
</table>

**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. 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. 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
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)

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 O4 Fe1 +++).

14. **HELIx**

**Cols. 1-6 HELIX**

8-10 Serial number (helix number)
12-14 Helix identifier (right justified)(i)
16-18 Residue name
20 Chain identifier
22-25 Residue seq. no. Initial residue of helix(ii)
26 Insertion code
28-30 Residue name
32 Chain identifier
34-37 Residue seq. no. Terminal residue of helix
38 Insertion code
39-40 Class of helix(iii)
41-70 Comment

**FORMAT** (6A1,1X,I3,1X,A3,2(I1X,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\textsubscript{10} 6 Left-handed α
7 Left-handed ω 8 Left-handed γ
9 27 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
18-20 Residue name
22 Chain identifier
23-26 Residue seq. no.
27 Insertion code
29-31 Residue name
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
46-48 Residue name
50 Chain identifier
51-54 Residue seq. no.
55 Insertion code
57-60 Atom name
61-63 Residue name
65 Chain identifier
66-69 Residue seq. no.
70 Insertion code


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. 1-5 TURN

8-10 Sequence number (Turn number)
12-14 Turn identifier (3 characters)
16-18 Residue name

\{
20 Chain identifier \quad Residue \ i
21-24 Residue seq. no.
25 Insertion code

\}
27-29 Residue name

\{
31 Chain identifier \quad Residue \ i+3 \\
\quad \text{(or \ i+2 \ for \ γ \ bend)}
32-35 Residue seq. no.
36 Insertion code

41-70 Comment

FORMAT (6A1,1X,13,1X,A3,1X,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. 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 seq. no.
36 Insertion code
41-70 Comment

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

18. **SITE**  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
23 Chain identifier  \{ First residue comprising site \\
24-27 Residue seq. no. \\
28 Insertion code \\
30-32 Residue name \\
34 Chain identifier  \{ Second residue comprising site \\
35-38 Residue seq. no. \\
39 Insertion code \\
41-43 Residue name \\
45 Chain identifier  \{ Third residue comprising site \\
46-49 Residue seq. no. \\
50 Insertion code \\
52-54 Residue name \\
56 Chain identifier  \{ Fourth residue comprising site \\

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.  CRST1  Cols. 1-6    CRST1

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  0_{11}  0_{12}  0_{13}  T_1
ORIGX2  0_{21}  0_{22}  0_{23}  T_2
ORIGX3  0_{31}  0_{32}  0_{33}  T_3

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

Note: Let the original submitted co-ordinates be X_{sub}, Y_{sub}, Z_{sub} and the orthogonal Å coordinates contained in the file be X, Y, Z. Then

\[ X_{sub} = 0_{11}X + 0_{12}Y + 0_{13}Z + T_1 \]
\[ Y_{sub} = 0_{21}X + 0_{22}Y + 0_{23}Z + T_2 \]
\[ Z_{sub} = 0_{31}X + 0_{32}Y + 0_{33}Z + T_3 \]
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**

<table>
<thead>
<tr>
<th></th>
<th>Cols. 1-6</th>
<th>11-20</th>
<th>21-30</th>
<th>31-40</th>
<th>46-55</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCALE1</td>
<td>S₁₁</td>
<td>S₁₂</td>
<td>S₁₃</td>
<td>U₁</td>
<td></td>
</tr>
<tr>
<td>SCALE2</td>
<td>S₂₁</td>
<td>S₂₂</td>
<td>S₂₃</td>
<td>U₂</td>
<td></td>
</tr>
<tr>
<td>SCALE3</td>
<td>S₃₁</td>
<td>S₃₂</td>
<td>S₃₃</td>
<td>U₃</td>
<td></td>
</tr>
</tbody>
</table>

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

**Note:** Let the orthogonal Å coordinates be X,Y,Z. Let the fractional cell coordinates be Xfrac, Yfrac, Zfrac. Then

\[
\begin{align*}
X_{\text{frac}} &= S_{11}X + S_{12}Y + S_{13}Z + U_1 \\
Y_{\text{frac}} &= S_{21}X + S_{22}Y + S_{23}Z + U_2 \\
Z_{\text{frac}} &= S_{31}X + S_{32}Y + S_{33}Z + U_3
\end{align*}
\]

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 \( \hat{a},\hat{b},\hat{c} \) are unit vectors in the orthogonal Å system and \( \hat{a},\hat{b},\hat{c} \) are unit vectors in the crystallographic system, then:

(i) the \( \hat{a},\hat{b},\hat{c} \), and \( \hat{a},\hat{b},\hat{c} \), systems have the same origin

(ii) \( \hat{a} \) is parallel to \( \hat{a} \)

(iii) \( \hat{b} \) is parallel to \( \hat{c} \times \hat{a} \)

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

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

<table>
<thead>
<tr>
<th>MTRIX1</th>
<th>Ser.no.</th>
<th>M₁₁</th>
<th>M₁₂</th>
<th>M₁₃</th>
<th>V₁</th>
<th>IGIVEN</th>
</tr>
</thead>
<tbody>
<tr>
<td>MTRIX2</td>
<td>Ser.no.</td>
<td>M₂₁</td>
<td>M₂₂</td>
<td>M₂₃</td>
<td>V₂</td>
<td>IGIVEN</td>
</tr>
<tr>
<td>MTRIX3</td>
<td>Ser.no.</td>
<td>M₃₁</td>
<td>M₃₂</td>
<td>M₃₃</td>
<td>V₃</td>
<td>IGIVEN</td>
</tr>
</tbody>
</table>

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 components of

21-30 translation vector

31-40

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

<table>
<thead>
<tr>
<th>Cols. 1-4 ATOM or 1-6 HETATM</th>
</tr>
</thead>
</table>

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 A 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,I13)

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

\[
U(eq) = \frac{1}{3}(U(1,1) + U(2,2) + U(33) + U(1,2)a*b*abcos\gamma + U(1,3)a*c*accos\beta + U(2,3)b*c*bccos\alpha)
\]

where \((a,b,c)\) are the direct cell dimensions given in the CRISTAL record and \((a^*,b^*,c^*)\) are the reciprocal cell dimensions.
25. **SIGATM**

Col. 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**

Col. 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

\[
U(eq) = \frac{1}{3}(U(1,1) + U(2,2) + U(3,3) + 2 \cdot U(1,2) \cdot a \cdot b \cdot \cos \gamma + 2 \cdot U(1,3) \cdot a \cdot c \cdot \cos \beta + 2 \cdot U(2,3) \cdot b \cdot c \cdot \cos \alpha)
\]

where \((a, b, c)\) are the direct cell dimensions given in the CRYSOL 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. MASTER 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
36-40 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. \textbf{END} \hfill End-of-entry record

Cols. 1-3 \textbf{END}

FORMAT (6A1)
APPENDIX A - COORDINATE SYSTEMS AND TRANSFORMATIONS

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

If \( \mathbf{\hat{a}}, \mathbf{\hat{b}}, \) and \( \mathbf{\hat{c}} \) describe the crystallographic cell edges and \( \mathbf{\hat{A}}, \mathbf{\hat{B}}, \) and \( \mathbf{\hat{C}} \) are unit vectors in the default orthogonal \( \mathbf{\hat{A}} \) system, then

a. \( \mathbf{\hat{A}}, \mathbf{\hat{B}}, \mathbf{\hat{C}} \) and \( \mathbf{\hat{a}}, \mathbf{\hat{b}}, \mathbf{\hat{c}} \) have the same origin.

b. \( \mathbf{\hat{A}} \) is parallel to \( \mathbf{\hat{a}} \).

c. \( \mathbf{\hat{B}} \) is parallel to \( \mathbf{\hat{c}} \times \mathbf{\hat{A}} \).

d. \( \mathbf{\hat{C}} \) is parallel to \( \mathbf{\hat{a}} \times \mathbf{\hat{b}} \) (i.e., \( \mathbf{\hat{c}}^* \)).

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

\[
\begin{bmatrix}
a & b \cos \gamma & c \cos \beta \\
0 & b \sin \gamma & c(\cos \alpha \cos \beta \cos \gamma)/\sin \gamma \\
0 & 0 & V/(ab \sin \gamma)
\end{bmatrix}
\]

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., \( \mathbf{x}_{\text{sub}} \)

(ii) A transformation from \( \mathbf{x}_{\text{sub}} \) to the orthogonal \( \mathbf{\hat{A}} \) coordinates stored in the Data Bank \((\mathbf{\hat{x}})\), i.e.,
\[0_{\text{sub}} \mathbf{x}_{\text{sub}} + \mathbf{t}_{\text{sub}} = \mathbf{\hat{x}}.\]
(iii) A transformation from $\hat{x}_{\text{sub}}$ to fractional crystallographic coordinates $\hat{x}_{\text{frac}}$

i.e., $S_{\text{sub}} \hat{x}_{\text{sub}} + U_{\text{sub}} = \hat{x}_{\text{frac}}$

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

i.e., $M_{\text{sub}} \hat{x}_{\text{sub}} + V_{\text{sub}} = \hat{x}_{\text{sub}}'$

Note: The notation $\hat{x}_{\text{sub}}$ is used for the column vector $X_{\text{sub}}, Y_{\text{sub}}, Z_{\text{sub}}$, 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 Angstroms ($\hat{x}$)

$$\hat{x} = O_{\text{sub}} \hat{x}_{\text{sub}} + \hat{t}_{\text{sub}}$$

(ii) The ORIGX transformation from stored to original coordinates ($0, \hat{t}$).

From above

$$\hat{x} = O_{\text{sub}} \hat{x}_{\text{sub}} + \hat{t}_{\text{sub}}$$

whence

$$O_{\text{sub}} \hat{x}_{\text{sub}} = \hat{x} - \hat{t}_{\text{sub}}$$

$$\therefore$$

$$\hat{x}_{\text{sub}} = O^{-1}_{\text{sub}} \hat{x} + (-O^{-1}_{\text{sub}} \hat{t}_{\text{sub}})$$

Thus

$$0 = O^{-1}_{\text{sub}}$$

and

$$\hat{t} = -O^{-1}_{\text{sub}} \hat{t}_{\text{sub}}$$
(iii) The SCALE transformation from stored to fractional coordinates \((S, \hat{\imath})\).

From above

\[ \hat{\imath}_{\text{frac}} = S_{\text{sub}} \hat{\imath}_{\text{sub}} + \hat{\imath}_{\text{sub}} \]

but

\[ \hat{\imath}_{\text{sub}} = O^{-1}_{\text{sub}} \hat{\imath}_{\text{sub}} + (-O^{-1}_{\text{sub}} \hat{\imath}_{\text{sub}}) \]

\[ \therefore \quad \hat{\imath}_{\text{frac}} = S_{\text{sub}} O^{-1}_{\text{sub}} \hat{\imath}_{\text{sub}} + (-S_{\text{sub}} O^{-1}_{\text{sub}} \hat{\imath}_{\text{sub}}) + \hat{\imath}_{\text{sub}} \]

i.e.,

\[ \hat{\imath}_{\text{frac}} = S_{\text{sub}} O^{-1}_{\text{sub}} \hat{\imath}_{\text{sub}} + (-S_{\text{sub}} O^{-1}_{\text{sub}} \hat{\imath}_{\text{sub}}) + \hat{\imath}_{\text{sub}} \]

\[ \therefore \quad S_{\text{sub}} = S_{\text{sub}} O^{-1}_{\text{sub}} \]

and

\[ \hat{\imath}_{\text{sub}} = -(S_{\text{sub}} O^{-1}_{\text{sub}} \hat{\imath}_{\text{sub}}) + \hat{\imath}_{\text{sub}} \]

(iv) The MTRIX transformation(s) expressing non-crystallographic symmetry in the space of the stored coordinates \((M, \hat{\jmath})\).

\[ \hat{\imath}'_{\text{sub}} = M_{\text{sub}} \hat{\imath}'_{\text{sub}} + \hat{\jmath}_{\text{sub}} \]

\[ \hat{\imath}' = O_{\text{sub}} \hat{\imath}'_{\text{sub}} + \hat{\jmath}_{\text{sub}} \]

\[ = O_{\text{sub}} \{M_{\text{sub}} \hat{\imath}'_{\text{sub}} + \hat{\jmath}_{\text{sub}}\} + \hat{\jmath}_{\text{sub}} \]

but

\[ \hat{\imath}'_{\text{sub}} = O^{-1}_{\text{sub}} \hat{\imath}'_{\text{sub}} + (-O^{-1}_{\text{sub}} \hat{\imath}'_{\text{sub}}) \]

\[ \therefore \quad \hat{\imath}' = O_{\text{sub}} \{M_{\text{sub}} O^{-1}_{\text{sub}} \hat{\imath}'_{\text{sub}} + (-O^{-1}_{\text{sub}} \hat{\imath}'_{\text{sub}})\} + \hat{\jmath}_{\text{sub}} \]

\[ \therefore \quad M_{\text{sub}} = O_{\text{sub}} M_{\text{sub}} O^{-1}_{\text{sub}} \]

and

\[ \hat{\jmath}_{\text{sub}} = -O_{\text{sub}} M_{\text{sub}} O^{-1}_{\text{sub}} \hat{\jmath}_{\text{sub}} + O_{\text{sub}} \hat{\jmath}_{\text{sub}} \]

In summary the stored coordinates and transformations are:

\[ \hat{\imath} \quad \text{(ATOM, HETATM records)} \]

\[ O, \hat{\jmath} \quad \text{(ORIGX records)} \]

\[ S, \hat{\imath} \quad \text{(SCALE records)} \]

\[ M, \hat{\jmath} \quad \text{(MTRIX records)} \]
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: \(\alpha-A, \beta-B, \gamma-G, \delta-D, \varepsilon-E, \zeta-Z, \eta-H\)

(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.


The 1974 recommendations on the "Nomenclature of \(\alpha\)-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
E. 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.

G. 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:

<table>
<thead>
<tr>
<th>ATP</th>
<th>Coenzyme A</th>
<th>Flavin mononucleotide (FMN)</th>
<th>Heme</th>
<th>Methotrexate</th>
<th>NAD</th>
</tr>
</thead>
<tbody>
<tr>
<td>26</td>
<td>27</td>
<td>28</td>
<td>29</td>
<td>30</td>
<td>31</td>
</tr>
</tbody>
</table>
ATOM NAMES AND ORDER INDICATORS FOR THE COMMON RIBONUCLEOTIDES (SUPERSCRIBPTS DENOTE ORDINALS)
PROTEIN DATA BANK STANDARD NOMENCLATURE FOR ADENOSINE TRIPHOSPHATE (ATP)

\[ C_{10}H_{12}N_5O_{13}P_3 \]

(1) Order indicators are given as preceding superscripts.
Protein Data Bank Standard Nomenclature for Coenzyme A

C_{21}H_{36}N_{7}O_{16}P_{3}S

(i) Order indicators are given as preceding superscripts.
PROTEIN DATA BANK STANDARD NOMENCLATURE FOR FLAVIN MONONUCLEOTIDE (FMN)

C_{17}H_{21}N_{4}O_{9}P

(1) Order indicators are given as preceding superscripts.
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 METHOTREXATE
\[ C_{20}H_{22}N_8O_5 \]  
(1) Order indicators are given as preceding superscripts.
PROTEIN DATA BANK STANDARD NOMENCLATURE FOR NICOTINAMIDE ADENINE DINUCLEOTIDE (NAD)

(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

<table>
<thead>
<tr>
<th>Residue</th>
<th>Abbr.</th>
<th>Synonym</th>
<th>Residue</th>
<th>Abbr.</th>
<th>Synonym</th>
</tr>
</thead>
<tbody>
<tr>
<td>γ-Aminobutyric Acid</td>
<td>ABU</td>
<td>Homoserine</td>
<td>HSE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acidic unknown</td>
<td>ACD</td>
<td>Hydroxylysine</td>
<td>HYL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alanine</td>
<td>ALA</td>
<td>Hydroxyproline</td>
<td>HYP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>β-Alanine</td>
<td>ALB</td>
<td>Isoleucine</td>
<td>ILE</td>
<td></td>
<td>I,ILU</td>
</tr>
<tr>
<td>Aliphatic unknown</td>
<td>ALI</td>
<td>Leucine</td>
<td>LEU</td>
<td>L</td>
<td></td>
</tr>
<tr>
<td>Arginine</td>
<td>ARG</td>
<td>Lysine</td>
<td>LYS</td>
<td>K</td>
<td></td>
</tr>
<tr>
<td>Aromatic unknown</td>
<td>ARF</td>
<td>Methionine</td>
<td>MET</td>
<td>M</td>
<td></td>
</tr>
<tr>
<td>Asparagine</td>
<td>ASN</td>
<td>Ornithine</td>
<td>ΨRN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aspartic acid</td>
<td>ASP</td>
<td>Pyrrolidone carboxylic acid</td>
<td>PCA</td>
<td>PGA</td>
<td></td>
</tr>
<tr>
<td>ASP/ASN ambiguous</td>
<td>ASX</td>
<td>Phenylalanine</td>
<td>PHE</td>
<td>F</td>
<td></td>
</tr>
<tr>
<td>Basic unknown</td>
<td>BAS</td>
<td>Proline</td>
<td>PRO</td>
<td>P,PRO,PRZ</td>
<td></td>
</tr>
<tr>
<td>Betaine</td>
<td>BET</td>
<td>Sarcosine</td>
<td>SAR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cysteine</td>
<td>CYS</td>
<td>Serine</td>
<td>SER</td>
<td>S</td>
<td></td>
</tr>
<tr>
<td>Cystine</td>
<td>CYS</td>
<td>Taurine</td>
<td>TAU</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glutamine</td>
<td>GLN</td>
<td>Threonine</td>
<td>THR</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td>Glutamic acid</td>
<td>GLU</td>
<td>Thryoxine</td>
<td>THY</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GLU/GLN Ambiguous</td>
<td>GLX</td>
<td>Tryptophan</td>
<td>TRP</td>
<td>W,TRY</td>
<td></td>
</tr>
<tr>
<td>Glycine</td>
<td>GLY</td>
<td>Tyrosine</td>
<td>TYR</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>Histidine</td>
<td>HIS</td>
<td>Valine</td>
<td>VAL</td>
<td>V</td>
<td></td>
</tr>
</tbody>
</table>
Notes:


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 "☐" 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.

<table>
<thead>
<tr>
<th>Residue</th>
<th>Abbr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adenosine</td>
<td>A</td>
</tr>
<tr>
<td>1-Methyladenosine</td>
<td>1MA</td>
</tr>
<tr>
<td>Cytidine</td>
<td>C</td>
</tr>
<tr>
<td>5-Methylcytidine</td>
<td>5MC</td>
</tr>
<tr>
<td>2'-O-Methylcytidine</td>
<td>☐MC</td>
</tr>
<tr>
<td>Guanosine</td>
<td>G</td>
</tr>
<tr>
<td>1-Methylguanosine</td>
<td>1MG</td>
</tr>
<tr>
<td>N(2)-Methylguanosine</td>
<td>2MG</td>
</tr>
<tr>
<td>N(2)-Dimethylguanosine</td>
<td>M2G</td>
</tr>
<tr>
<td>7-Methylguanosine</td>
<td>7MG</td>
</tr>
<tr>
<td>2'-O-Methylguanosine</td>
<td>☐MG</td>
</tr>
<tr>
<td>Wybutosine</td>
<td>YG</td>
</tr>
<tr>
<td>Inosine</td>
<td>I</td>
</tr>
<tr>
<td>Thymidine</td>
<td>T</td>
</tr>
<tr>
<td>Uridine</td>
<td>U</td>
</tr>
<tr>
<td>Modified Uridine</td>
<td>+U</td>
</tr>
<tr>
<td>Dihydouridine</td>
<td>H2U</td>
</tr>
<tr>
<td>Ribosylthymidine</td>
<td>5MU</td>
</tr>
<tr>
<td>Pseudouridine</td>
<td>PSU</td>
</tr>
</tbody>
</table>

Note: To avoid confusion here within residue abbreviations, the alphabetic character is written "☐" and the numeric "0". This convention is not observed elsewhere throughout these specifications.
C. Miscellaneous

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

<table>
<thead>
<tr>
<th>Residue</th>
<th>Abbr.</th>
<th>Synonym</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acetyl</td>
<td>ACE</td>
<td></td>
</tr>
<tr>
<td>Formyl</td>
<td>FOR</td>
<td></td>
</tr>
<tr>
<td>Water</td>
<td>HOH</td>
<td>H2O, WAT, H2O</td>
</tr>
<tr>
<td>Unknown</td>
<td>UNK</td>
<td></td>
</tr>
</tbody>
</table>

Note: To avoid confusion here within residue abbreviations, the alphabetic character is written "Φ" 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. 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=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 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. **COMPD.** 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.

3. **SOURCE.** 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.

4. **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.

5. **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 termini (e.g., during an activation process) are not included. Residues excised from the chain (not at the termini, 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 'O' and cols. 20-22 contain 'UNK'.

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:

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

In this sketch each line represents a continuous 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. CRYSCL. 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. **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.

19. **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.

20. **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.

21. **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.).

22. **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 work 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.

<table>
<thead>
<tr>
<th>Name</th>
<th>Code</th>
<th>Formula</th>
<th>Mol. wt.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alanine</td>
<td>ALA</td>
<td>C₃ H₇ N₁ O₂</td>
<td>89.09</td>
</tr>
<tr>
<td>Arginine</td>
<td>ARG</td>
<td>C₆ H₁₄ N₄ O₂</td>
<td>174.20</td>
</tr>
<tr>
<td>Asparagine</td>
<td>ASN</td>
<td>C₄ H₈ N₂ O₂</td>
<td>132.12</td>
</tr>
<tr>
<td>Aspartic acid</td>
<td>ASP</td>
<td>C₄ H₇ N₁ O₄</td>
<td>133.10</td>
</tr>
<tr>
<td>ASP/ASN ambiguous</td>
<td>ASX</td>
<td>C₄ H₇ ¹/₂ N₁ ²/₄ O₃³/₄</td>
<td>132.61</td>
</tr>
<tr>
<td>Cysteine</td>
<td>CYS</td>
<td>C₃ H₇ N₁ O₂ S₁</td>
<td>121.15</td>
</tr>
<tr>
<td>Glutamine</td>
<td>GLN</td>
<td>C₅ H₁₀ N₂ O₃</td>
<td>146.15</td>
</tr>
<tr>
<td>Glutamic acid</td>
<td>GLU</td>
<td>C₅ H₉ ¹/₂ N₁ ²/₄ O₄³/₄</td>
<td>147.13</td>
</tr>
<tr>
<td>GLU/GLN ambiguous</td>
<td>GLX</td>
<td>C₅ H₉ N₁ O₃</td>
<td>146.64</td>
</tr>
<tr>
<td>Glycine</td>
<td>GLY</td>
<td>C₂ H₅ N₁ O₂</td>
<td>75.07</td>
</tr>
<tr>
<td>Histidine</td>
<td>HIS</td>
<td>C₆ H₉ N₃ O₂</td>
<td>155.16</td>
</tr>
<tr>
<td>Isoleucine</td>
<td>ILE</td>
<td>C₆ H₁₃ N₁ O₂</td>
<td>131.17</td>
</tr>
<tr>
<td>Leucine</td>
<td>LEU</td>
<td>C₆ H₁₃ N₁ O₂</td>
<td>131.17</td>
</tr>
<tr>
<td>Lysine</td>
<td>LYS</td>
<td>C₆ H₁₄ N₂ O₂</td>
<td>146.19</td>
</tr>
<tr>
<td>Methionine</td>
<td>MET</td>
<td>C₅ H₁₁ N₁ O₂ S₁</td>
<td>149.21</td>
</tr>
<tr>
<td>Phenylalanine</td>
<td>PHE</td>
<td>C₉ H₁₁ N₁ O₂</td>
<td>165.19</td>
</tr>
<tr>
<td>Proline</td>
<td>PRO</td>
<td>C₅ H₉ N₁ O₂</td>
<td>115.13</td>
</tr>
<tr>
<td>Serine</td>
<td>SER</td>
<td>C₃ H₇ N₁ O₃</td>
<td>105.09</td>
</tr>
<tr>
<td>Threonine</td>
<td>THR</td>
<td>C₄ H₉ N₁ O₃</td>
<td>119.12</td>
</tr>
<tr>
<td>Tryptophan</td>
<td>TRP</td>
<td>C₁₁ H₁₂ N₂ O₂</td>
<td>204.23</td>
</tr>
<tr>
<td>Tyrosine</td>
<td>TYR</td>
<td>C₉ H₁₁ N₁ O₃</td>
<td>181.19</td>
</tr>
<tr>
<td>Valine</td>
<td>VAL</td>
<td>C₅ H₁₁ N₁ O₂</td>
<td>117.15</td>
</tr>
<tr>
<td>Undetermined</td>
<td>UNK</td>
<td>C₅ H₆ N₁ O₃ (approx.)</td>
<td>128.16</td>
</tr>
<tr>
<td>Name</td>
<td>Code</td>
<td>Formula</td>
<td>Mol. wt.</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>------</td>
<td>----------------</td>
<td>----------</td>
</tr>
<tr>
<td>Nucleotides</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adenosine</td>
<td>A</td>
<td>C10 H14 N5 O7 P1</td>
<td>347.22</td>
</tr>
<tr>
<td>1-Methyladenosine</td>
<td>1MA</td>
<td>C11 H16 N5 O7 P1</td>
<td>361.25</td>
</tr>
<tr>
<td>Cytidine</td>
<td>C</td>
<td>C9 H14 N3 O8 P1</td>
<td>323.20</td>
</tr>
<tr>
<td>5-Methylcytidine</td>
<td>5MC</td>
<td>C10 H16 N3 O8 P1</td>
<td>337.23</td>
</tr>
<tr>
<td>2'-O-Methylcytidine</td>
<td>φMC</td>
<td>C10 H17 N3 O8 P1</td>
<td>338.23</td>
</tr>
<tr>
<td>Guanosine</td>
<td>G</td>
<td>C10 H14 N5 O8 P1</td>
<td>363.22</td>
</tr>
<tr>
<td>1-Methylguanosine</td>
<td>1MG</td>
<td>C11 H16 N5 O8 P1</td>
<td>377.25</td>
</tr>
<tr>
<td>N(2)-Methylguanosine</td>
<td>2MG</td>
<td>C11 H16 N5 O8 P1</td>
<td>377.25</td>
</tr>
<tr>
<td>N(2)-Dimethylguanosine</td>
<td>M2G</td>
<td>C12 H18 N5 O8 P1</td>
<td>391.28</td>
</tr>
<tr>
<td>7-Methylguanosine</td>
<td>7MG</td>
<td>C11 H10 N5 O8 P1</td>
<td>377.25</td>
</tr>
<tr>
<td>2'-O-Methylguanosine</td>
<td>φMG</td>
<td>C11 H16 N5 O8 P1</td>
<td>377.25</td>
</tr>
<tr>
<td>Wybutosine</td>
<td>YG</td>
<td>C21 H26 N6 O11 P1</td>
<td>587.48</td>
</tr>
<tr>
<td>Inosine</td>
<td>I</td>
<td>C10 H13 N4 O8 P1</td>
<td>348.21</td>
</tr>
<tr>
<td>Thymidine</td>
<td>T</td>
<td>C10 H15 N2 O8 P1</td>
<td>322.21</td>
</tr>
<tr>
<td>Uridine</td>
<td>U</td>
<td>C9 H13 N2 O9 P1</td>
<td>324.18</td>
</tr>
<tr>
<td>Dihydouridine</td>
<td>H2U</td>
<td>C9 H15 N2 O9 P1</td>
<td>326.20</td>
</tr>
<tr>
<td>Ribosylthymidine</td>
<td>5MU</td>
<td>C10 H16 N2 O10 P1</td>
<td>355.22</td>
</tr>
<tr>
<td>Pseudouridine</td>
<td>PSU</td>
<td>C9 H13 N2 O9 P1</td>
<td>324.18</td>
</tr>
<tr>
<td>Miscellaneous</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acetic Acid</td>
<td>ACE</td>
<td>C2 H4 O2</td>
<td>60.05</td>
</tr>
<tr>
<td>Formic Acid</td>
<td>FOR</td>
<td>C1 H2 O2</td>
<td>46.03</td>
</tr>
<tr>
<td>Water</td>
<td>HOH</td>
<td>H2 O1</td>
<td>18.015</td>
</tr>
</tbody>
</table>