PROTEIN DATA BANK

FILE RECORD FORMATS

July 1976

MACROMOLECULAR STRUCTURE FILE

For each parameter set the file consists of records each of 80 characters.

The record sequence is as follows:-

HEADER	:	Date entered into data bank; identification code.
COMPND	:	Name of protein and identifying information
SOURCE	:	Species from which protein has been obtained
AUTHOR	:	Names of contributor and coauthors
JRNL	:	Literature citation which defines coordinate set
REMARK	:	General remarks
SEQRES	:	The residue sequence
FTNOTE	:	Footnotes relating to specific atoms or residues
HET	:	Identification of non-standard groups
HELIX	:	Identification of helical substructures
SHEET	:	Identification of sheet substructures
TURN	:	Identification of hairpin turns (β bends)
SITE	:	Identification of groups comprising the various sites
CRYST1	:	Unit cell data
ORIGX	:	Transformation matrix (orthogonal Å coords → submitted coords)
SCALE	:	Transformation matrix (orthogonal Å coords → fractional crystallographic coords)
MTRIX	:	Transformation matrix (expression of non-crystallographic symmetry)
ATOM	:	Atomic coordinate records
HETATM	:	Coordinates for heterogenous components
TER	:	Chain terminator
CONECT	:	Connectivity records
MASTER	:	Master control record with checksums of total number of records in the file, broken down by record type
END	:	End of data record

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

RECORD FORMATS

1. HEADER Cols. 1-6 HEADER

   11-50 Reserved for future use

   51-60 Date entered into data bank

   63-70 Identification code

   FORMAT (6A1,4X,60A1)

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

2. COMPND Cols. 1-6 COMPND

   11-70 Name of molecule

   FORMAT (6A1,4X,60A1)

3. SOURCE Cols. 1-6 SOURCE

   11-70 Species from which molecule has been obtained

   FORMAT (6A1,4X,60A1)

4. AUTHOR Cols. 1-6 AUTHOR

   11-70 Names of contributor and coauthors

   FORMAT (6A1,4X,60A1)

5. JRNL Cols. 1-4 JRNL

   11-70 Literature citation which defines the coordinate set

   FORMAT (6A1,4X,60A1)

6. REMARK Cols. 1-6 REMARK

   8-10 Remark number

   12-70 Remarks

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

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

7. SEQRES Cols. 1-6 SEQRES

   9-10 Serial number

   12 Chain identifier

   14-17 No. 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))

   Serial numbers are reset to 1 for each new chain.

8. FTNOTE Cols. 1-6 FTNOTE

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

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

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).

9. HET

Cols.	1-3	HET
	8-10	Heterogen identifier
	12-18	Sequence identifier (see page 8)
	21-25	No of atoms in heterogenous group
	31-70	Text

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

10. HELIX

Cols.	1-6	HELIX	
	8-10	Serial number (Helix number)	
	12-14	Helix identifier (right justified)	
	16-18	Residue name	Initial residue of helix
	20	Chain identifier	Initial residue of helix
	22-25	Residue seq. no.	Initial residue of helix
	26	Insertion code	Initial residue of helix
	28-30	Residue name	Terminal residue of helix
	32	Chain identifier	Terminal residue of helix
	34-37	Residue seq. no.	Terminal residue of helix
	38	Insertion code	Terminal residue of helix
	39-40	Class of helix (Appendix A)	
	41-70	Remarks	

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

Notes: (i) The initial residue has a lower sequence number than the terminal residue.

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

11. <u>SHEET</u>

Cols.		
1-5	SHEET	
8-10	Strand number(i)(v)	
12-14	Sheet identifier(i)	(right justified)
15-16	Number of strands	
18-20	Residue name	Initial residue
22	Chain identifier	Initial residue
23-26	Residue seq. no.(ii)	Initial residue
27	Insertion code	Initial residue
29-31	Residue name	Terminal residue
33	Chain identifier	Terminal residue
34-37	Residue seq. no.	Terminal residue
38	Insertion code	Terminal residue
39-40	Sense of above strand with respect to previous strand(iii)	
42-45	Atom name	Registration(iv)
46-48	Residue name	Registration(iv)
50	Chain identifier	Registration(iv)
51-54	Residue seq. no.	Registration(iv)
55	Insertion code	Registration(iv)
57-60	Atom name	Registration(iv)
61-63	Residue name	Registration(iv)
65	Chain identifier	Registration(iv)
66-69	Residue seq. no.	Registration(iv)
70	Insertion code	Registration(iv)

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

<u>Notes:</u>

(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.

12. <u>TURN</u> Cols.

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

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

<u>Note</u>: - These records identify the hairpin turns (β bends) in the structure which do not occur in helices.

13. <u>SITE</u> Cols.

1-4	SITE	
8-10	Sequence no.(i)	
12-14	Site identifier(ii) (right justified)	
16-17	No. of residues comprising site(iii)	
19-21	Residue name	First residue comprising site
23	Chain identifier	
24-27	Residue seq. no.	
28	Insertion code	
30-32	Residue name	Second residue comprising site
34	Chain identifier	
35-38	Residue seq. no.	
39	Insertion code	
41-43	Residue name	Third residue comprising site
45	Chain identifier	
46-49	Residue seq. no.	
50	Insertion code	
52-54	Residue name	Fourth residue comprising site
55	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))

<u>Notes</u>:
(i) Sequence numbers are reset to 1 for each new site.
(ii) Site identifiers should be fully explained in the REMARKs.
(iii) If a site is comprised of more than four residues then these may be specified on additional records bearing the same site identifier.

14. 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)

15. 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 coordinates be $X_oY_oZ_o$

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

Then
$$X_o = O_{11}X + O_{12}Y + O_{13}Z + T_1$$
$$Y_o = O_{21}X + O_{22}Y + O_{23}Z + T_2$$
$$Z_o = O_{31}X + O_{32}Y + O_{33}Z + T_3$$

Even if this is an identity transformation (unit matrix, null vecto it is supplied. See below under SCALE for a definition of the default orthogonal Å system. Appendix B details the derivation of this coordinate transformation.

16. 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 x y z

Then
$$x = S_{11}X + S_{12}Y + S_{13}Z + U_1$$
$$y = S_{21}X + S_{22}Y + S_{23}Z + U_2$$
$$z = S_{31}X + S_{32}Y + S_{33}Z + U_3$$

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

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

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

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

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

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

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

Appendix B details the derivation of this coordinate transformation

17. MATRIX

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

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

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

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

$$X' = M_{11}X + M_{12}Y + M_{13}Z + V_1$$
$$Y' = M_{21}X + M_{22}Y + M_{23}Z + V_2$$
$$Z' = M_{31}X + M_{32}Y + M_{33}Z + V_3$$

The serial number is constant for each group of three MTRIX records. More than one group of MTRIX records may be present to describe different local symmetry operations.

Appendix B details the derivation of this coordinate transformation

18. ATOM   Atomic coordinate records

Cols.	1-4	ATOM
	7-11	Atom serial number (i)
	13-16	Atom name (ii)
	17	Alternate location indicator (iii)
	18-20	Residue name (iv)
	21-27	Sequence identifier (v)
	31-38	X } Orthogonal Å coordinates
	39-46	Y } Orthogonal Å coordinates
	47-54	Z } Orthogonal Å coordinates
	55-60	Occupancy
	61-66	Temperature factor
	68-70	Footnote number

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

Note: -
(i) Atoms occur in order of their serial numbers which always increase starting from the N-terminal residue. Within each amino acid residue the atoms are ordered as indicated in Appendix C. If the residue sequence is known certain serial numbers may be omitted to allow for future insertion of any missing atoms. If the sequence is not reliably known these serial numbers are simply ordinals.
(ii) Appendix C
(iii) Alternate locations for atoms may be denoted by A, B, C, etc. here.
(iv) Standard residue names are given in Appendix D, other components are defined in HET records.
(v) The sequence identifier would be a composite field made up as follows:

Cols.	21	Reserved for future expansion
	22	Chain identifier, e.g. A for hemoglobin α chain
	23-26	Residue sequence number
	27	Code for insertions of residues, e.g. 66A,66B,etc

19. HETATOM   Cols. 1-6 HETATM
7-70 as for ATOM records

20. TER   Cols. 1-3 TER
7-11 Serial number
18-20 Residue name
21-27 Sequence identifier

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. The residue defined on these TER records is the carboxy-terminal residue of the chain in question.

21. <u>CONECT</u> Connectivity records

Cols.		
1-6	CONECT	
7-11	Serial number	
12-16	Covalent bond connectivity (serial number of bonded atoms)	
17-21		
22-26		
27-31		
32-36	Hydrogen bond connectivity (Donor)	
37-41		
42-46		
47-51	Hydrogen bond connectivity (Acceptor)	
52-56		
57-61		

FORMAT (6A1,11I5)

Note: - Serial numbers are identical to those in cols 7-11 of the ATOM/HETATM records and connectivity entries correspond to these serial numbers. For metal atoms, which may form more than four covalent bonds, a second CONECT record with the same serial number may be used. Either all or none of the connectivity of an atom must be specified.

22. <u>MASTER</u> Cols.

Cols.	
1-6	MASTER
11-15	Number of REMARK records
16-20	Number of FTNOTE records
21-25	Number of HET records
26-30	Number of HELIX records
31-35	Number of SHEET records
36-40	Number of TURN records
41-45	Number of SITE records
46-50	Number of MTRIX records (ORIGX+SCALE+MTRIX)
51-55	Number of ATOM records (ATOM+HETATM)
56-60	Number of TER records
61-65	Number of CONECT records
66-70	Number of SEQRES records

FORMAT (6A1,5X,12I5)

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

23. <u>END</u> End of data record

Cols. 1-3 END

FORMAT (6A1)

APPENDIX A

Classification of Helical Substructures

Class	Substructure
1	Right-handed α
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

Default is class 1.

APPENDIX B

Coordinate Transformations

Usually the depositor will have supplied--

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

(ii) A transformation from $\vec{X}_o$ to the orthogonal Å coordinates stored in the data bank ($\vec{X}$)

i.e. $\underset{\sim}{O}_o \vec{X}_o + \vec{T}_o = \vec{X}$

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

The default $\underset{\sim}{O}_o, \vec{T}_o$ transformation which transforms fractional crystallographic coordinate into the default orthogonal Å system is

$$\underset{\sim}{O}_o = \begin{pmatrix} a & b\cos\gamma & c\cos\beta \\ 0 & b\sin\gamma & c(\cos\alpha - \cos\beta\cos\gamma)/\sin\gamma \\ 0 & 0 & V/(ab\sin\gamma) \end{pmatrix}, \quad \vec{T}_o = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

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

i.e. $\underset{\sim}{S}_o \vec{X}_o + \vec{U}_o = \vec{x}$

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

i.e. $\underset{\sim}{M}_o \vec{X}_o + \vec{V}_o = \vec{X}_o'$

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

The stored quantities are

(i) The coordinates in orthogonal Ångstroms ($\vec{X}$)

$$\vec{X} = \underset{\sim}{O}_o \vec{X}_o + \vec{T}_o$$

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

From above $$\vec{X} = \underset{\sim}{O}_o \vec{X}_o + \vec{T}_o$$

whence $$\underset{\sim}{O}_o \vec{X}_o = \vec{X} - \vec{T}_o$$

$$\therefore \quad \vec{X}_o = \underset{\sim}{O}_o^{-1} \vec{X} + (-\underset{\sim}{O}_o^{-1} \vec{T}_o)$$

Thus $$\underset{\sim}{O} = \underset{\sim}{O}_o^{-1}$$

and $$\vec{T} = -\underset{\sim}{O}_o^{-1} \vec{T}_o$$

(iii) The SCALE transformation from stored to fractional coordinates ($\underset{\sim}{S}$, $\vec{U}$).

From above $$\vec{x} = \underset{\sim}{S}_o \vec{X}_o + \vec{U}_o$$

but $$\vec{X}_o = \underset{\sim}{O}_o^{-1} \vec{X} + (-\underset{\sim}{O}_o^{-1} \vec{T}_o)$$

$$\therefore \quad \vec{x} = \underset{\sim}{S}_o [\underset{\sim}{O}_o^{-1} \vec{X} + (-\underset{\sim}{O}_o^{-1} \vec{T}_o] + \vec{U}_o$$

i.e. $$\vec{x} = \underset{\sim}{S}_o \underset{\sim}{O}_o^{-1} \vec{X} + (-\underset{\sim}{S}_o \underset{\sim}{O}_o^{-1} \vec{T}_o) + \vec{U}_o$$

$$\therefore \quad \underset{\sim}{S} = \underset{\sim}{S}_o \underset{\sim}{O}_o^{-1}$$

and $$\vec{U} = -(\underset{\sim}{S}_o \underset{\sim}{O}_o^{-1} \vec{T}_o) + \vec{U}_o$$

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

$$\vec{X}_o' = \underset{\sim}{M}_o \vec{X}_o + \vec{V}_o$$

$$\therefore \quad \vec{X}' = \underset{\sim}{O}_o \vec{X}_o' + \vec{T}_o$$

$$= \underset{\sim}{O}_o \left\{ \underset{\sim}{M}_o \vec{X}_o + V_o \right\} + \vec{T}_o$$

but $$\vec{X}_o = \underset{\sim}{O}_o^{-1} \vec{X} + (-\underset{\sim}{O}_o^{-1} T_o)$$

$$\therefore \quad \vec{X}' = \underset{\sim}{O}_o \left\{ \underset{\sim}{M}_o [\underset{\sim}{O}_o^{-1} \vec{X} + (-\underset{\sim}{O}_o^{-1} T_o)] + \vec{V}_o \right\} + \vec{T}_o$$

$$\therefore \quad \underset{\sim}{M} = \underset{\sim}{O}_o \underset{\sim}{M}_o \underset{\sim}{O}_o^{-1}$$

and $$\vec{V} = -\underset{\sim}{O}_o \underset{\sim}{M}_o \underset{\sim}{O}_o^{-1} \vec{T}_o + \underset{\sim}{O}_o \vec{V}_o + \vec{T}_o$$

In summary the stored coordinates and transformations are:

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

APPENDIX C

Atom Identifiers

Atom names, remoteness codes and order indicators for the common amino acids.

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

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

α-A	δ-D	η-H
β-B	ε-E	
γ-G	ζ-Z	

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

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

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

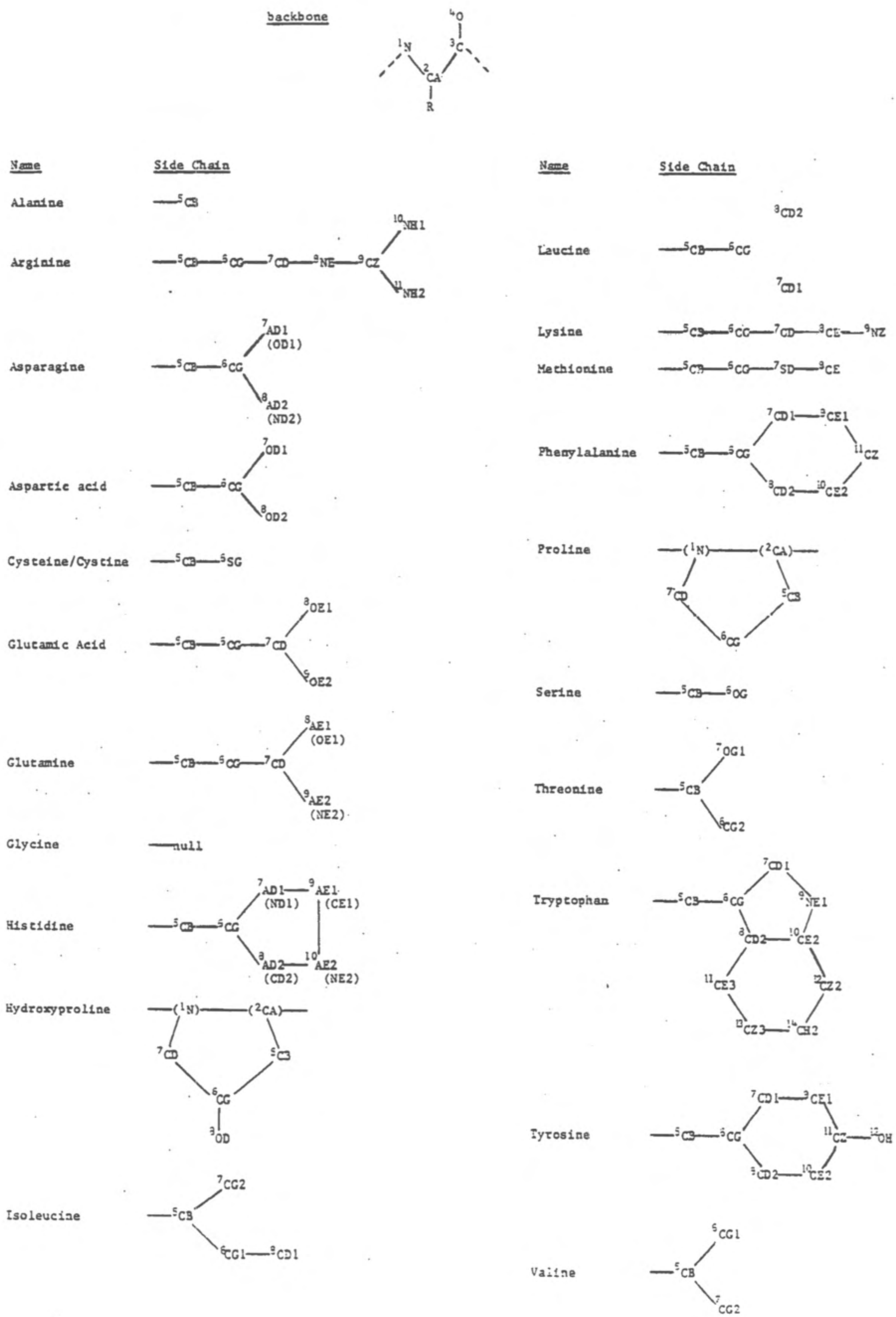
Four characters are reserved for these atom names--they are assigned as follows:

1-2	Chemical symbol - right justified
3	Remoteness indicator (alphabetic)
4	Branch designator (numeric)

---

[1]IUPAC-IUB Commission on Biochemical Nomenclature. "Abbreviations and Symbols for the Description of the Conformation of Polypeptide Chains. Tentative Rules (1969)," Biochemistry, 9, 3471 (1970).
The 1974 recommendations on the "Nomenclature of α-Amino Acids" (Biochemistry, 14, 449 (1975)) provides a scheme based on normal rules for organic compounds but this scheme will not be used here.

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



APPENDIX D

Residue Names and Abbreviations

Residue	Abb.	Syn.	Residue	Abb.	Syn.
Acidic unknown	ACD		Homoserine	HSE	
Acetyl	ACE		Hydroxyproline	HYP	
Alanine	ALA		Hydroxylysine	HYL	
β-Alanine	ALB		Isoleucine	ILE	
Aliphatic unknown	ALI		Leucine	LEU	
γ-Aminobutyric acid	ABU		Lysine	LYS	
Arginine	ARG		Methionine	MET	
Aromatic unknown	ARO		Ornithine	ORN	
Asparagine	ASN		Phenylalanine	PHE	
Aspartic acid	ASP		Proline	PRO	
ASP/ASN ambiguous	ASX		Pyrollidone carboxylic acid	PCA	
Basic unknown	BAS		Sarcosine	SAR	
Betaine	BET		Serine	SER	
Cysteine	CYS	CYH,CSH	Taurine	TAU	
Cystine	CYS	CSS	Terminator	TER	
Formyl	FOR		Threonine	THR	
Glutamic acid	GLU		Thyroxine	THY	
Glutamine	GLN		Tryptophan	TRP	TRY
GLU/GLN ambiguous	GLX		Tyrosine	TYR	
Glycine	GLY		Unknown	UNK	
Heterogen	HET		Valine	VAL	
Histidine	HIS		Water	HOH	H2O,WAT

Note: - Standard residue abbreviations conform to the IUPAC-IUB rules in J. Biol. Chem., 241, 527, 2491 (1966).
Recognizable synonyms, such as those above, will be changed to the standard abbreviation.