INTERNATIONAL UNION OF
PURE AND APPLIED CHEMISTRY
AND
INTERNATIONAL UNION OF BIOCHEMISTRY

A ONE-LETTER NOTATION FOR AMINO ACID SEQUENCES (DEFINITIVE RULES)

Issued by the
IUPAC–IUB Commission on Biochemical Nomenclature
1971

LONDON
BUTTERWORTHS
639

IUPAC–IUB COMMISSION
ON BIOCHEMICAL NOMENCLATURE†

A ONE-LETTER NOTATION FOR AMINO ACID SEQUENCES

In 1968 the IUPAC–IUB Commission on Biochemical Nomenclature (CBN) published in *IUPAC Information Bulletin No. 32* tentative rules for the use of a one-letter notation for amino acid sequences. These definitive rules are based on the comments received.

1. GENERAL CONSIDERATIONS

Various difficulties are encountered when presenting the formulas of long protein sequences in the usual three-letter symbols[1a]. Space is often at a premium. A one-letter code minimizes this difficulty and has other distinct advantages. In summarizing large amounts of data or in the alignment of homologous protein sequences, it is important that the patterns in the sequences be condensed and simplified as much as possible. Computer techniques are increasingly applied for the storage of sequences of hundreds of amino acid residues and for their evaluation. For these purposes, a one-letter code is the best solution. Finally, a one-letter code is useful in the labeling of individual amino-acid sidechains in three-dimensional pictures of protein molecules.

The possibility of using one-letter symbols was mentioned by Gamow and Yčas[2] in 1958. The idea was systematized by Šorm *et al.*[3] in 1961. It was used by this group[4–10] and also by Fitch[11] in several papers on the structure of proteins. In extensive compilations of protein structures, Eck and Dayhoff (see refs 12 to 14) systematically used one-letter symbols derived partly from the code of Šorm and Keil. Independent proposals were made by Wiswesser[15] and by Braunstein‡.

In view of the increasing number of different notations and the attending problems, the IUPAC–IUB Commission on Biochemical Nomenclature (CBN) undertook the task of drafting a single notation for one-letter symbols. The present proposal was evolved by a CBN subcommission (composed

† Those who have served on the Commission for varying periods during 1967–71 are the following. Present Members are shown by an asterisk *. O. Hoffmann-Ostenhof* (Chairman), W. E. Cohn* (Secretary), A. E. Braunstein*, J. S. Fruton, B. L. Horecker*, P. Karlson*, B. Keil*, W. Klyne*, C. Liébecq*, E. C. Slater, E. C. Webb*, W. J. Whelan*.

Comments on and suggestions for future revisions of these rules should be sent to: Prof. O. Hoffmann-Ostenhof, Lehrkanzel für Biochemie der Universität Wien, Währinger Strasse 38, A-1090 Wien, Austria.

‡ A. E. Braunstein, personal proposal to CBN.

of B. Keil, R. V. Eck, M. O. Dayhoff and W. E. Cohn); it is based principally on the system evolved by Dayhoff and Eck[12–14].

2. LIMITS OF APPLICATION

In publications, CBN recommends that one-letter symbols be used only in comparisons of long sequences in tables, lists or figures, and for such special use as tagging three-dimensional models of proteins. They should not be used in simple text nor for original reports of experimental details of sequences. This system is not suitable for reporting the details of peptide synthesis, for example, where a fuller description of substituents is needed and where uncommon amino acids may occur. It should not be used in papers where the single-letter system for nucleoside sequences is employed (see ref. 1b, Section N-3.2), as in representing codons, etc.

3. PRINCIPLES OF THE ONE-LETTER CODE

3.1 The letter written at the left-hand end is that of the amino acid residue carrying the free amino group and the letter written at the right-hand end is that of the amino acid residue carrying the free carboxyl group. The absence of punctuation beyond either end of a sequence implies that it is known to be the amino or carboxyl end of the protein. A fragmentary sequence is to be preceded or followed by a slash (/) to indicate that it is not known to be the end of the complete protein (see 'Comment' in Section 8.2).

3.2 Initial letters are used where there is no ambiguity. There are six such cases: cysteine, histidine, isoleucine, methionine, serine and valine. All the other amino acids share the initial letters A, G, L, P or T; therefore, assignments of them must be somewhat arbitrary. These letters are assigned to the most frequently occurring and structurally most simple amino acids. On this basis, the letters A, G, L, P and T are assigned to alanine, glycine, leucine, proline and threonine, respectively.

3.3 The assignment of the other abbreviations is more arbitrary. However, certain clues are helpful. Two are phonetically suggestive, F for *ph*enylalanine, and R for *ar*ginine. For tryptophan, the double ring in the molecule is associated with the bulky letter W. The letters N and Q are assigned to asparagine and glutamine, respectively; D and E are assigned to aspartic acid and glutamic acid, respectively. This leaves lysine and tyrosine, to which K and Y are assigned. These are chosen rather than any of the few other remaining letters because they are alphabetically near the initial letters L and T. U and O are avoided because U is easily confused with V in handwritten work and O is confused with G, Q, C and D in imperfect computer print-outs and also with zero. J is avoided for linguistic reasons.

3.4 Two other abbreviations are necessary in order to avoid ambiguity. B is assigned to aspartic acid or asparagine when this distinction has not been determined. Z is assigned when glutamic acid and glutamine have not been distinguished. X means that the identity of an amino acid is undetermined or that the amino acid is atypical.

4. ABBREVIATIONS (IN ALPHABETICAL ORDER)

Table 1

Amino acid	1-letter symbol	Amino acid	1-letter symbol
*A*lanine	A	*M*ethionine	M
*A*rginine	R	*Ph*enylalanine	F
Asparagine	N } B*	*P*roline	P
Aspartic acid	D }	*S*erine	S
*C*ysteine	C	*T*hreonine	T
Glutamine	Q } Z†	*T*ryptophan	W
Glutamic acid	E }	*T*yrosine	Y
*G*lycine	G	*V*aline	V
*H*istidine	H		
*I*soleucine	I	Unknown or	
*L*eucine	L	'other'	X
Lysine	K		

Table 2

1-letter symbol	3-letter symbol	1-letter symbol	3-letter symbol
A	Ala	N	Asn
B*	Asx	P	Pro
C	Cys	Q	Gln
D	Asp	R	Arg
E	Glu	S	Ser
F	Phe	T	Thr
G	Gly	V	Val
H	His	W	Trp
I	Ile	Y	Tyr
K	Lys	Z†	Glx
L	Leu	X	—
M	Met		

* For 'aspartic acid or asparagine'. † For 'glutamic acid or glutamine'.

5. SPACING

A very important use of the one-letter notation is in presenting alignments of many homologous sequences. In printing, it often happens that the alignment is not perfectly maintained because of the variable size of the letters and the variable amount of punctuation. This effect can be very troublesome in extensive comparisons. Therefore, **a single typewriter space is left between letters, either as a blank or occupied by punctuation** (see Sections 6, 7 and 8). The alignment is preserved by allowing **exactly the same spacing for each letter, each blank, and each punctuation mark,** as in typewritten material or, if printed, as in 'typewriter type fount'.

6. KNOWN AND UNKNOWN SEQUENCES

A **blank** between letters indicates that the sequence is **known.** (See also 'Comment' in Section 8.2.) As in the three-letter notation, **parentheses** and **commas** are used to indicate regions in which the sequence is **unknown** or **undetermined.**

Example

In three-letter symbols:

Ser–Tyr–Cys–Phe–His(Asn, Gln)Cys(Pro, Val)Lys–Gly

In one-letter symbols:

S Y C F H(N, Q)C(P, V)K G

7. JUXTAPOSITION OF UNKNOWN SEQUENCES KNOWN TO BE CONNECTED

Consider the two sequences, one completely known, the other containing peptides of unknown internal sequence.

(a) Ala–Cys–Asp–Glu–Phe–Gly–His–Ile–Lys–Leu–Met–Asn–Pro–Gln

(b) (Ala, Cys, Asp)(Arg, Ser) (Gly, His, Ile) Lys–Leu–Met–Asn–Pro–Gln

In one-letter notation, these become:

(a) A C D E F G H I K L M N P Q

(b) (A,C,D)(R,S)(G,H,I)K L M N P Q

↑ ↑

In the second illustration, two punctuation marks have been crowded into each of two single spaces (indicated by the arrows). In a computer print-out, this would not be possible. A single one-space symbol must be used. Here **= is used for)(to indicate the end of one unknown sequence and the beginning of another,** as shown below.

(a) A C D E F G H I K L M N P Q

(b) (A,C,D = R,S = G,H,I)K L M N P Q

↑ ↑

8. JUXTAPOSITION OF RESIDUES INFERRED, BUT NOT KNOWN, TO BE CONNECTED

Consider the following case in which peptides from a second sequence (d) can be aligned with a known, related sequence (c).

(c) A C D E F G H I K L M N P Q

(d) (A.C.D = R,S = G.H.I K L M N P Q

8.1 In this illustration, the sequences of two of the fragments (A.C.D and G.H.I in d), while not determined, are **inferred** with good confidence, which is indicated by **periods** instead of commas between their residues. Where such inferences can **not** be made with confidence, commas, which retain their original connotation of 'unknown sequence' (Section 6), should be used as in the R, S dipeptide.

8.2 The two **internal slashes (/) separate adjacent amino acids that come from different peptides not proven experimentally to be connected.** The third (end) slash indicates that Q is not experimentally proven to be at the carboxyl end of the protein, although it is at the carboxyl end of the P-Q dipeptidyl residue.

Comment—The absence of punctuation at the beginning or end of a complete polypeptide or protein sequence indicates the known amino or carboxyl terminal, respectively (see Section 3.1).

8.3 Depending on the experimental details and the nature of the inferences to be represented, even more elaborate punctuation may sometimes be required. It is essential, however, that **only one character (or a blank space of similar size) appear between the single letters** to preserve the spacing that is essential for comparisons (see Section 5).

REFERENCES

[1] IUPAC–IUB Tentative Rules: (a) *J. Biol. Chem.* **247**, 977 (1972); (b) *ibid* **245**, 5171 (1970); and elsewhere.

[2] G. Gamov and M. Yčas, *Symposium on Information Theory in Biology*, Pergamon: New York (1958).

[3] F, Šorm, B. Keil, J. Vaněček, V. Tomášek, O. Mikeš, B. Meloun, V. Kostka and V. Holeyšovský, *Collect. Czech. Chem. Commun.* **26**, 531 (1961).

[4] O. Mikeš, V. Holeyšovský, V. Tomášek, B. Keil and F. Šorm, *Collect. Czech. Chem. Commun.* **27**, 1964 (1962).

[5] V. Holeyšovský, B. Alexijev, V. Tomášek, O. Mikeš and F. Šorm, *Collect. Czech. Chem. Commun.* **27**, 2662 (1962).

[6] F. Šorm and B. Keil, *Advanc. Protein Chem.* **17**, 1967 (1962).

[7] O. Mikeš, V. Holeyšovský, V. Tomášek and F. Šorm, *Abstracts Sixth International Congress of Biochemistry, 1964, IUB Vol. 32*, p 169, II-136.

[8] B. Keil, Z. Prusík and F. Šorm, *Biochim. Biophys. Acta*, **78**, 559 (1963).

[9] O. Mikeš, Z. Prusík and F. Svoboda, *Collect. Czech. Chem. Commun.* **29**, 1193 (1964).

[10] F. Šorm, V. Holeyšovský, O. Mikeš and V. Tomášek, *Collect. Czech. Chem. Commun.* **30**, 2103 (1965).

[11] W. M. Fitch. *J. Molec. Biol.* **16**, 1, 9, 17 (1966).

[12] M. O. Dayhoff, R. V. Eck, M. A. Chang and M. R. Sochard, *Atlas of Protein Sequence and Structure*, National Biomedical Research Foundation: Silver Spring, Md (1965).

[13] R. V. Eck and M. O. Dayhoff, *Atlas of Protein Sequence and Structure*, National Biomedical Research Foundation: Silver Spring Md (1966).

[14] M. O. Dayhoff and R. V. Eck, *Atlas of Protein Sequence and Structure*, National Biomedical Research Foundation, Silver Spring, Md (1967–1968).

[15] W. J. Wiswesser, *Chem. Eng. News.* **42**, 4 (1964).