Biochemistry

 Volume 7, Number 8 August 12, 1968

IUPAC–IUB Commission on Biochemical Nomenclature. A One-Letter Notation for Amino Acid Sequences. Tentative Rules*

EDITORIAL NOTE: These tentative rules are printed for the benefit of those workers who have occasion to use one-letter abbreviations for amino acid residues in the course of their own work. As a rule, *Biochemistry will not accept* the use of these abbreviations in manuscripts that are to be published in this journal. Acceptable abbreviations have been formulated by the Commission on Biochemical Nomenclature and may be found in the CBN reports published earlier (*Biochemistry 5*, 2485 (1966); *6*, 362 (1967); *7*, 483 (1968)).

1. *General Considerations*

Various difficulties are encountered when presenting the formulas of long protein sequences in the usual three-letter symbols (IUPAC–IUB, 1966–1968). 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 this purpose, a one-letter code is the best solution. Finally, a one-letter code is useful in the labeling of individual amino acid side chains in three-dimensional pictures of protein molecules.

The possibility of using one-letter symbols was mentioned by Gamow and Yčas in 1958. The idea was systematized by Šorm *et al.* in 1961. It was used by this group (Mikes *et al.*, 1962, 1964a,b; Holeyšovský *et al.*, 1962; Keil *et al.*, 1963; Šorm and Keil, 1962; Šorm *et al.*, 1965) and also by Fitch (1966) in several papers on the structure of proteins. In extensive compilations of protein structures, Eck and Dayhoff (Dayhoff *et al.*, 1965; Dayhoff and Eck, 1968; Eck and Dayhoff, 1966) systematically used one-letter symbols derived partly from the code of Šorm and Keil. Independent proposals were made by Wiswesser (1964) and by Braunstein.[1]

In view of the increasing number of different notations and the attending problems, the IUPAC–IUB Commission on Biochemical Nomenclature (CBN) has undertaken the task of drafting a single notation for one-letter symbols. The present proposal was evolved by a CBN subcommission (composed of B. Keil, R. V. Eck, M. O. Dayhoff, and W. E. Cohn); it is based principally

* Document of the IUPAC–IUB Commission on Biochemical Nomenclature (CBN), approved by CBN in March 1968 and published by permission of the International Union of Pure and Applied Chemistry, the International Union of Biochemistry, and the official publishers to the International Union of Pure and Applied Chemistry, Messrs. Butterworths Scientific Publications. Comments on these Tentative Rules may be sent to any member of CBN: O. Hoffmann-Ostenhof (Chairman), W. E. Cohn (Secretary), A. E. Braunstein, J. S. Fruton, P. Karlson, B. Keil, W. Klyne, C. Liébecq, E. C. Slater, E. C. Webb, or corresponding member, N. Tamiya. Reprints of these Tentative Rules may be obtained from Waldo E. Cohn, Director, NAS–NRC Office of Biochemical Nomenclature, Oak Ridge National Laboratory, Box Y, Oak Ridge, Tenn. 37830.

[1] A. E. Braunstein, personal proposal to CBN.

on the most recent summary published by Dayhoff and Eck (1968).

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 or 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 (IUPAC–IUB, 1965, sections 5.4 and 5.5), 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, and assignments of them must therefore 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 bulky letter W. The letters N and Q are assigned to asparagine and glutamine, respectively; D and E are assigned to aspartic acid and gluamic 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 nearest 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 the amino acid is atypical.

4. *Abbreviations, in Alphabetical Orders*

Amino Acid		Abbreviation
*A*lanine	A	A Ala
*Ar*ginine	R	B Asx[a]
Asparagine	N > B[a]	C Cys
Aspartic acid	D > B[a]	D Asp
*C*ysteine	C	E Glu
Glutamine	Q > Z[b]	F Phe
Glutamic acid	E > Z[b]	G Gly
*G*lycine	G	H His
*H*istidine	H	I Ile
*I*soleucine	I	K Lys
*L*eucine	L	L Leu
Lysine	K	M Met
*M*ethionine	M	N Asn
*Ph*enylalanine	F	P Pro
*P*roline	P	Q Gln
*S*erine	S	R Arg
*T*hreonine	T	S Ser
Tryptophan	W	T Thr
T*y*rosine	Y	V Val
*V*aline	V	W Trp
Unknown or "other"	X	X Unknown or "other"
		Y Tyr
		Z Glx[b]

[a] For Asp or Asn (*i.e.*, for Asx). [b] For Glu or Gln (*i.e.*, for Glx).

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 6–8). 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 font."

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 (β-corticotropin releasing factor (Schally and Bowers, 1964)

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 printout, 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 *dots* 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

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

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

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

Fitch, W. M. (1966), *J. Mol. Biol. 16*, 1, 9, 17.

Gamow, G., and Yčas, M. (1958), Symposium on Information Theory in Biology, New York, N. Y., Pergamon.

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

IUPAC–IUB Tentative Rules, 1965 (1966), *Biochemistry, 5*, 1445.

IUPAC–IUB Tentative Rules (1966), *Biochemistry 5*, 2485; (1967), *Biochemistry 6*, 362; (1968), *Biochemistry 7*, 483.

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

Mikes, O., Holeyšovský, V., Tomášek, V., Keil, B., and Šorm, F. (1962), *Collection Czech. Chem. Commun. 27*, 1964.

Mikeš, O., Holeyšovský, V., Tomášek, V., and Šorm, F. (1964b), *6th Intern. Congr. Biochem., New York*, Abstr. II-136, p 169.

Mikeš, O., Prusík, Z., and Svoboda, F. (1964a), *Collection Czech. Chem. Commun. 29*, 1193.

Schally, A. V., and Bowers, C. Y. (1964), *Metabolism 13*, 1190.

Šorm, F., Holeyšovský, V., Mikeš, O., and Tomášek, V. (1965), *Collection Czech. Chem. Commun. 30*, 2103.

Šorm, F., and Keil, B. (1962), *Advan. Protein Chem. 17*, 1967.

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

Wiswesser, W. J. (1964), *Chem. Eng. News 42*, No. 20, 4.