

One-Letter Notation for Calculating Molecular Formulas and Searching Long-Chain Peptides in the *Index Chemicus Registry System**†

GABRIELLE S. REVESZ, CHARLES E. GRANITO, and EUGENE GARFIELD
Institute for Scientific Information, 325 Chestnut St., Philadelphia, Pa. 19106

Received February 3, 1970

The one-letter nomenclature for amino acids as proposed by the IUPAC-IUB Commission for Biochemical Nomenclature has been modified and expanded to include some less common amino acids and 39 commonly used substituents. A computer program has been developed to calculate the molecular formulas of long-chain peptides from these expanded notations. All new peptide sequences appearing in the literature are stored on magnetic tape in the *Index Chemicus Registry System (ICRS)* combining the Wiswesser Line Notation (for protecting and ester groups) with the IC-IUPAC one-letter code. These symbols have also been permuted by computer to produce an index which permits rapid look-up of all peptides containing any specified amino acid sequence.

It has been only within the last 20 years that stepped up research in the structure elucidation of long-chain polypeptides has made an abbreviated nomenclature for their amino acid building blocks essential. In his historical paper (1945), Brand¹ writes: "It will become imperative in the future to find a compact way of presenting the empirical formula of a protein in terms of its amino acid residues. We propose to use as symbols, whenever feasible, the first three letters of the amino acids: e.g., glycine residue, Gly; arginine residue, Arg; isoleucine residue, Ileu; etc.; the residues of cysteine, halfcystine, asparagine, and glutamine we propose to designate by CySH, (Cy S-), (Asp-NH₂), and (Glu-NH₂), respectively. Such a formula for β -lactoglobulin is given in the summary. We hope that a satisfactory convention will be gradually developed."

This proposal gained wide acceptance during the ensuing years and became the official nomenclature accepted by the IUPAC-IUB Commission on Biochemical Nomenclature, undergoing only minor changes during the next 20 years. Biochemists learned to associate these three-letter codes with protein structures rather than the real chemical structures of these amino acid residues. In fact, frequently, when actual chemical structures for a short peptide sequence are published, many a researcher is hard put to identify or name the amino acid residues. For large molecules—e.g., the Tobacco Mosaic Virus Coat Protein—drawing of structures would be a near impossibility, taking up reams and reams of paper and immeasurably valuable time. Assuming one would be willing to spend the time, energy, and money to do this, the result would be a confusing conglomeration of atomic symbols. Should a researcher now try to define clearly differences between the Beta-Haemoglobin of the human,

the sheep, gorilla, llama, horse, pig, etc., he would probably get lost in all his drawings.

It was not until 1958 that Gamov and Ycas² came up with the suggestion of using one-letter symbols instead of three-letter codes, and not until 1961 that this suggestion was systematized by Sorm *et al.*³ In 1964 Wiswesser⁴ published a proposition to use one-letter notations for long-chain peptides, and Eck and Dayhof^{5,6,7} published their "Atlas" for protein structure using such abbreviated notations.

By 1967, the IUPAC-IUB Commission on Biochemical Nomenclature had its final draft for "Tentative Rules on a One-Letter Notation for Amino Acid Sequences" and this was published in several biochemical journals in 1968. Heralding the publication of these "Tentative Rules" was an editorial in *Nature*,⁸ denouncing such a one-letter coding system, in which Maddox claimed that a chain of amino acids consisting of the sequence:

Ser-Ileu-Leu-Leu-Tyr

would spell:

S I L L Y

and that was exactly what he thought of the whole system. What he did not realize was the fact that computers have come to stay and can be effectively used for both calculating molecular formulas and searching for long-chain sequences.

About the same time Maddox published his editorial in *Nature*, the Institute for Scientific Information came out with its new *Index Chemicus Registry System (ICRS)*.⁹

In this system the structures of new compounds appearing in the chemical literature, and thus indexed in *Index Chemicus (IC)*,¹⁰ are translated into Wiswesser Line Notations (WLN).¹¹ These notations are then key-punched and sorted by computer to bring structurally similar compounds together for both structure and sub-structure searching. Since new amino acid sequences are

* Presented before the Division of Biological Chemistry, 158th Meeting, ACS, New York, Sept. 10, 1969.

† *Index Chemicus (IC)* and *Index Chemicus Registry System (ICRS)* are registered trade marks of the Institute for Scientific Information.

CALCULATING MOLECULAR FORMULAS OF LONG-CHAIN PEPTIDES

part of the *IC*, the problem of encoding these structures had to be considered.

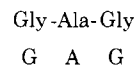
It was soon found impractical to encode amino acid sequences following the same rules used for all other organic compounds¹² because of the size of the resulting notations (Figure 1). Yet, during the past few years, the number of natural peptides whose structures have been elucidated and the number of new synthetic amino acid sequences have grown to such an extent that manual searching for these sequences has become prohibitive. The solution to the problem was found in the one-letter notation.

MODIFIED IC-IUPAC CODE

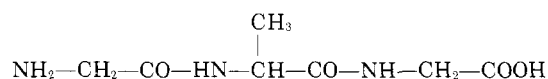
In September 1968, as a solution for the *ICRS* system, a modified *IC-IUPAC* code was proposed and published in *Nature*¹³ (Figure 2). The letters J, O, U, and X were added for Cit, Orn, Cys-Cys, and HyPro, respectively. It was further proposed that the system be open-ended and additional letter combinations be used for some rarely occurring amino acids.

To simplify computer programming, all code letters denote amino acids with one molecule of water removed, contrary to the IUPAC proposition wherein the first letter

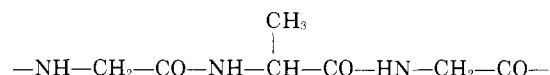
in the sequence represents an amino acid with its NH_2 group complete, while the last letter includes the OH needed to complete the terminal COOH group. Thus the sequence,



according to the IUPAC proposition means



while in the *ICRS* system it means



Since the H at the beginning and the OH at the end of the sequence are missing, they have to be accounted for. Furthermore, the amino group is frequently protected, and the acid group is frequently changed to an ester or amide. Since the *ICRS* is basically a register of compounds according to the WLN, the notation has been used to indicate these end groups. In its simplest form, the above sequence would be found in the *ICRS* as



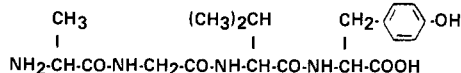
H meaning Hydrogen and Q meaning OH in WLN. The three slashes indicate that between two such sets is an open-ended amino acid chain. Outside of the three slashes, conventional WLN symbols are used. Thus, the same amino acid sequence, protected by benzyloxycarbonyl and esterified to give the methyl ester, would be represented by the following notation:



where R1OV = $\text{PhCH}_2\text{OC}(\text{O})$ and O1 = OCH_3 (Figure 3).

Handling of substituents on the amino acids within a given chain was to be the next problem. A study has been made of the most commonly occurring substituents starting with $\text{CH}_3\text{C}(\text{O})$ (Acetyl), and a numerical code, denoted within parentheses, has been given to 39 of them

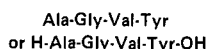
STRUCTURAL DIAGRAM



WISWESSER LINE NOTATION



CONVENTIONAL PEPTIDE NOMENCLATURE



IC ONE LETTER NOTATION



Figure 1. Various representations of a peptide

A	Ala	L	Leu	W	Trp
B	Asx	M	Met	X	HyPro
C	CySH	N	Asn	Y	Tyr
D	Asp	O	Orn	Z	Glx
E	Glu	P	Pro	A@	Cysteic acid
F	Phe	Q	Gln	B@	β -NH ₂ -butyric acid
G	Gly	R	Arg	C@	Homoserine
H	His	S	Ser	D@	Pyro-glutamic acid
I	Ile	T	Thr	E@	β -NH ₂ -valeric acid
J	Cit	U	Cys-SCy	F@	Sarcosine
K	Lys	V	Val	G@	β -alanine

Figure 2. IC-IUPAC amino acid code key

A	alkyl	T	heterocyclic ring
E	bromine	U	double bond
G	chlorine	V	carbonyl
J	halogen	W	nonlinear dioxo
K	quaternary nitrogen	X	quaternary carbon
L	carbocyclic ring	Y	tertiary carbon
M	secondary nitrogen	Z	amino or amido
N	tertiary nitrogen atom	&	punctuation mark
Q	hydroxyl	-	separator or connective
R	benzene ring	/	multiplier stop

Numerals preceded by a space are multipliers of preceding notation symbols; or within ring signs L . . . J or T . . . J show the number of multicyclic points in the ring structure. Numerals not preceded by a space show ring sizes if within the ring signs; elsewhere numerals show the length of internally saturated, unbranched alkyl chains and segments. Letters following a space and hyphen are proposed as symbols with special meanings to denote stereoisomerism.

All international atomic symbols except K, U, V, W, Y, Cl, and Br are used.

Figure 3. WLN symbols

(Figure 4). Thus, acetyl becomes (01), tosyl (02), benzyl (03), etc. A simple sort by computer will now take all compounds between a set of six slashes out from all other organic compounds and produce the Peptide Index of *ICRS*. This index, as it is produced today, contains some 2000 to 3000 new peptides each year (Figure 5).

The stage is now set for some more complex computer operations than simple sorts—e.g., rapid calculation of the molecular formulas for long-chain peptides.

COMPUTER CALCULATION OF MOLECULAR FORMULAS FOR PEPTIDES

Each letter code is assigned a molecular formula representing an amino acid minus water: thus, G = C₂H₅NO, A = C₃H₅NO, etc. Each substituent (numerical code in parenthesis) is also assigned a formula which represents the group in question minus a hydrogen—e.g., CH₃C(O) = Ac=(01) is coded as CH₂CO to take into account the hydrogen which has been replaced by the substituent. WLN's for protecting ester or amide groups are also assigned a molecular formula—e.g., R1OV = PhCH₂OCO = C₈H₇O₂; Z = NH₂; etc. This information is used to generate molecular formulas for amino acid chains up to 123 characters in length (Figure 6).

The program also takes into consideration cyclic peptides and peptides of more than one chain bonded together. Such sequences are coded by giving the sequence number of the bonded amino acids at the end of the notation.

ICRS SUBSTITUENT CODE KEY

01 acetyl	14 4-NO ₂ -phthaloyl	27 carbomethoxy
02 tosyl	15 tetra-Cl-phthaloyl	28 beta-aminoethyl
03 benzyl	16 penta-Cl-phenyl-O	29 succinimidooxy
04 4-NO ₂ -benzyl	17 2-NO ₂ -phenylsulfenyl	30 t-Bu
05 2, 4-di-NO ₂ -benzyl	18 2, 4, 5-tri-Cl-phenyl	31 formyl
06 1-adamantylloxycarbonyl	19 2, 4, 6-tri-Me-benzyl	32 NH ₂
07 t-BuO-CO	20 4-Br-benzylloxycarbonyl	33 Me
08 Cbo	21 4-NH ₂ -phenylacetyl	34 Et
09 NO ₂	22 4-NH ₂ -benzoyl	35 2,4,6-tri-NO ₂ -phenyl
10 phthalimido	23 decanoyl	36 SO ₃ H
11 benzhydryl	24 stearoyl	37 phthaloyl
12 trityl	25 cerotoyl	38 acetamidomethyl
13 t-amyl-OCO	26 tri-F-acetyl	39 benzoyl

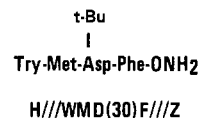
Figure 4. Code numbers used for substituents on the amino acids

PEPTIDES

	ABSTR. CPD. NO. NO.
H///VTLK///Q	112382-7
H///VYAC///Q	112166-43
H///YCQY///Q	33
R1OV///K(07)D(30)K(07)R///Z	112193-14
R1OV///PPK(07)D(30)K(07)R///Z . . .	8
R1OV///PYK(07)M///MZ	4
R1OV///PYK(07)M///O1	3
R1OV///PYK(07)ME(30)HFRWG///Q . .	7
R1OV///PYK(07)ME(30)HFRWGSPPK(07)D(30)K(07)R///Z	19
R1OV///SPPK(07)D(30)K(07)E///Z	9

Figure 5. Typical peptide list in the *Index Chemicus Registry System*

Molecular Formula Calculation

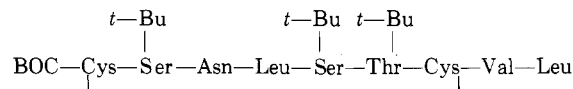


IC 114252-15

	C	H	N	O	S
H///:	—	1	—	—	—
W :	11	10	2	1	—
M :	5	9	1	1	1
D :	4	5	1	3	—
(30) :	4	8	—	—	—
F :	9	9	1	1	—
///Z :	—	2	1	—	—
MF =	C 33	H 44	N 6	O 6	S

Figure 6. Typical table for molecular formula calculation

Thus:



would be coded as: 1X&&OV///CS(30)NLS(30)T(30)CVL 1-7///Q where 1-7 signifies bonding between the first and seventh amino acid residues. The numbers appear without parentheses, follow a blank space, and are separated by hyphens. Instructions are contained in the program for subtracting one hydrogen for each number appearing at the end of the notation to account for the bonds between each two amino acids.

CHECKER PROGRAM

The program, which calculates the molecular formula for any given sequence, also checks the accuracy of molecular formulas which have been calculated manually or are given by an author in his original publication. This is done by calculating the molecular formula and comparing it to the existing one. Should there be a discrepancy between the two molecular formulas, an error message is printed out, thus assuring accuracy of both the encoding and the correct molecular formula as printed in the *Index Chemicus* (Figure 7).

PERMUTED INDEXES FOR PEPTIDES (PIP)

The modified *IC-IUPAC* code may also be used for generating permuted indexes. The use of indexes of permuted WLN's^{14, 15} for substructure searching of compounds has already found wide acceptance. This approach has been expanded to cover the peptide codes.

Each amino acid symbol appearing in a new peptide leads to one entry in a permuted index for peptides. True WLN symbols (outside "///...///") and symbols within parentheses do not generate entries for this index.

As shown in Figure 8, the peptide BHA3 isolated from cod insulin has eight entries. The 580 new peptides

CALCULATING MOLECULAR FORMULAS OF LONG-CHAIN PEPTIDES

Checker Program

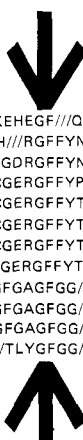
```

IC #      Cpd #      Encoded peptide
106237    7          1X&&OV///K(07)E(30)TAAP///MZ
C 040 H 071 N 009 O 013      GIVEN M.F.
C 040 H 071 N 009 O 013      CALCULATED M.F.

106237    8          H///K(07)PE(30)O(07)QH///O1
C 051 H 083 N 011 O 015      GIVEN M.F.
C 047 H 079 N 011 O 014      CALCULATED M.F.

••• M.F. IS NOT CORRECT
                                ERROR MESSAGE
  
```

Figure 7. Printout from the checker program



	IC NO.	CPD NO.
H///DEHKKEHEGF///QI	107332	11
H///RGFFYNPK///QI	109318	2
H///MAPPOHLC(28)GSHLVDALYLVC(28)GDRGFFYNPK///QI	109318	3
H///FVNQHLGGAHLVEALYLVCGERGFFYTPKA///QI	106363	3
H///NAHLGSHLVEALYLVCGERGFFYTPKA///QI	106362	5
H///FVNAHLGSHLVEALYLVCGERGFFYTPKA///QI	106362	2
H///FVAQLGSHLVEALYLVCGERGFFYTPKA///QI	106362	1
H///VNQHLGSHLVEALYLVCGERGFFYTPKA///QI	106362	4
R10V///AAGFGAGFGG///O1R B D F	107477	2
H///AAGFGAGFGG///I	107477	3
R10V///AAGFGAGFGG///O1R B D F	107477	2
H///TLYGFGG///QI	109040	3

Figure 9. Permuted peptide index

reported in *IC* during the first quarter of 1969 led to 6151 entries (10.6 per peptide). An alphabetic sort of all entries (disregarding numerical symbols within parentheses signifying substituents) is printed out as the index.

The value of this permuted index is that it permits searches on the basis of amino acid sequence. For example, Figure 9 shows part of the G section of the index for peptides noted in the first quarter of 1969. Peptide BHA3 (*IC* No. 109318-2), noted earlier, is represented on the second line. Even a casual inspection of Figure 9 reveals the close similarity between peptides from *IC* abstracts 106362 and 109318.

The peptides in 106362 were isolated from bovine insulin at the Institute for Physiology, University of Tubingen, Germany.¹⁶ Those in 109318 were isolated from cod insulin at the National Environment Research Council, University of Aberdeen, Torry, Scotland.¹⁷ The similarities between these peptides is striking but by no means an exception. Throughout the index, related peptides are brought together by amino acid sequence. Thus, one can quickly locate all new peptides which are structurally related.

The ability to locate new peptides solely on the basis of amino acid sequence facilitates literature searching for those interested in peptide research. Plans are now being finalized at the Institute for Scientific Information to make indexes of this type commercially available to the scientific community.

Entries for Peptide BHA3 (Arg-Gly-Phe-Phe-Tyr-Asn-Pro-Lys) *IC* 109318-2

```

      *
H///RGFFYNPK///Q
H///RGFFYNPK///Q
H///RGFFYNPK///Q
H///RGFFYNPK///Q
H///RGFFYNPK///Q
H///RGFFYNPK///Q
H///RGFFYNPK///Q
H///RGFFYNPK///Q
      *
  
```

* REFERENCE COLUMN IN INDEX

Figure 8. Permuted entries for the peptide BHA3

ACKNOWLEDGMENT

The authors wish to thank Jack E. Byk for the computer programming that was involved in this work and for his continuous help and encouragement.

LITERATURE CITED

- Brand, E., L. J. Saidel, W. H. Goldwater, Beatrice Kassell, F. J. Ryan, *J. Amer. Chem. Soc.* **67**, 1524-32 (1945).
- Gamov, G., and M. Ycas, "Symposium on Information Theory in Biology," Pergamon Press, New York (1958).
- Sorm, F., B. Keil, J. Vanecek, V. Tomasek, O. Mikes, B. Meloun, V. Kostka, and V. Holeysovsky, *Collect. Czech. Chem. Commun.* **26**, 531 (1961).
- Wiswesser, W. J., *Chem. Eng. News* **42**, 4 (1964).
- Dayhoff, M. O., R. V. Eck, M. A. Chang, and M. R. Sochard, "Atlas of Protein Sequence and Structure," Nat. Biomed. Res. Found. 1965.
- Eck, R. V., and M. O. Dayhoff, "Atlas of Protein Sequence and Structure," Nat. Biomed. Res. Found., 1966.
- Dayhoff, M. O., and R. V. Eck, "Atlas of Protein Sequence and Structure," Nat. Biomed. Res. Found., 1968.
- Editorial, *Nature* **218**, 10 (1968).
- Garfield E., G. S. Revesz, C. E. Granito, H. A. Dorr, M. M. Calderon, and A. Warner, "Index Chemicus Registry System—A Pragmatic Approach to Sub-Structure Chemical Retrieval," *J. CHEM. DOC.* **10**, 54-58 (1970).
- Revesz, G. S., and A. Warner, "Retrieving Chemical Information with Index Chemicus," *J. CHEM. DOC.* **2**, 106-9 (1969).
- Wiswesser, W. J., "A Line-Formula Chemical Notation," Thomas Y. Cromwell Co., New York (1954).
- Smith, E. G., "The Wiswesser Line-Formula Chemical Notation," McGraw-Hill, New York (1968).
- Revesz, G. S., "Simplified Notations for Peptides in Computer Compatible Format," *Nature* **219**, 1113 (1968).
- Sorter, P., C. E. Granito, J. C. Gilmer, Alan Gelberg, and E. A. Metcalf, *J. CHEM. DOC.* **4**, 56 (1964).
- Granito, C. E., J. E. Schultz, G. W. Gibson, Alan Gelberg, R. J. Williams, and E. A. Metcalf, *J. CHEM. DOC.* **5**, 229 (1965).
- Weber, U., and G. Weitzel, *Z. Physiol. Chem.* **349**, (20), 1431-3 (1968).
- Reid, K. B. M., and P. T. Grant, *Biochem. J.* **110** (2), 289-96 (1968).