BIOCHEMICAL JOURNAL LETTERS

Suggestion to research groups working on protein and peptide sequence

Today with the widespread use of computers an increasing number of people need to enter new published protein sequences to update or create their own data bank. After having typed in the sequence the operator is confronted with the time-consuming job of checking its accuracy.

We therefore propose that every report dealing with the publication or the revision of a polypeptide sequence show in addition to the sequence itself four parameters which will facilitate the detection of typographical and keyboard errors. The operator will be able to recalculate those parameters from the data he has entered, and by comparing them to those published along with the sequence he will instantly be aware of the validity of his entry.

(1) The first parameter, *COMP*, represents the amino-acid composition of the polypeptide by using the international one-letter notation for each amino-acid (IUPAC–IUB Commission on Biochemical Nomenclature, 1969; Table 1), each letter being followed by its integral frequency in the sequence. It is only necessary to include B, Z and X (Asx, Glx and Xaa) if those residues are present in the sequence, but all other amino-acid codes should be included even if their frequency is zero in the sequence.

(2) The second parameter, *NR*, is the total number of amino-acid residues in the sequence.

(3) The third parameter, *MMP*, is the molecular mass of the polypeptide calculated from the structure before it undergoes any post-translational modification. This value is computed from the sequence by using the residue masses tabulated by Hunt *et al.* (1978), which are reprinted in Table 1.

(4) The last parameter is a checking number, *CN*, defined as follows:

$$CN = \sum_{i=1}^{NR} AA_{(i)} \cdot i$$

Where $AA_{(i)}$ is the reference number (defined in Table 1) of the amino-acid residue found in the *i*th position.

Table 1. *Symbols, residue masses and reference numbers (AA) for the common amino acids*

Amino acid	One-letter notation	Residue mass (Da)	AA
Ala	A	89.09	1
Arg	R	174.20	2
Asn	N	132.12	3
Asp	D	133.10	4
Cys	C	121.15	5
Gln	Q	146.15	6
Glu	E	147.13	7
Gly	G	75.07	8
His	H	155.16	9
Ile	I	131.17	10
Leu	L	131.17	11
Lys	K	146.19	12
Met	M	149.21	13
Phe	F	165.19	14
Pro	P	115.13	15
Ser	S	105.09	16
Thr	T	119.12	17
Trp	W	204.23	18
Tyr	Y	181.19	19
Val	V	117.15	20
Asx	B	132.65	21
Glx	Z	146.64	22
Xaa	X	128.16	23
H_2O	–	18.015	–

The first three parameters will reveal accidental additions, deletions and substitutions. The parameter *CN*, on the other hand, is sensitive to internal inversions in the sequence, a type of error among the most likely to occur. The parameter *CN* should be a number sufficiently large so that the probability that two different sequences might give the same value is small. The probability will of course never be nil, as a unique numerical representation of a polypeptide sequence would be longer than its expression in the one-letter notation. We have therefore chosen to define *CN* in such a way as for it to be moderately simple to compute while allowing it to serve as a fairly sensitive indicator of keyboard errors.

The *CN* for a protein 100 residues long will vary between 10^4 and 10^5, while for 1000 residues the value can reach up to 10^7. An error in the *C*-terminal of the sequence will cause a much greater error in

Peptide: H E L P I H A T E M A T H

CN computation: $CN = 1\cdot9+2\cdot7+3\cdot11+4\cdot15+5\cdot10+6\cdot9+7\cdot1+8\cdot17+9\cdot7+10\cdot13+11\cdot1+12\cdot17+13\cdot9 = 788$

$COMP = A_2R_0N_0D_0C_0Q_0E_2G_0H_3I_1L_1K_0M_1F_0P_1S_0T_2W_0Y_0V_0$

$NR = 13 \quad MMP = 1186.66 \quad CN = 788$

Fig. 1. *Computation of parameters for an imaginary polypeptide*

the *CN* value than one in the *N*-terminal. This feature can be used to locate an entry error more quickly.

A practical example of the computation of the *CN* of a small imaginary polypeptide is given in Fig. 1, along with a proposition for the presentation of the four parameters.

I thank Dr. R. E. Offord for helpful discussions and comments on the manuscript, and the 'Fonds National Suisse de la Recherche Scientifique' for computing time.

Amos BAIROCH
Département de Biochimie Médicale,
Centre Médical Universitaire,
1, rue Michel-Servet,
CH-1211 Génève 4, Switzerland

(*Received 15 February 1982*)

Hunt, L. T., Barker, W. C. & Dayhoff, M. O. (1978) in *Atlas of Protein Sequence and Structure* (Dayhoff, M. O., ed.), vol. 5, suppl. 3, pp. 25–27, National Biomedical Research Foundation, Washington DC

IUPAC–IUB Commission on Biochemical Nomenclature (1969) *Biochem. J.* **113**, 1–4

0306-3275/82/050527-02$01.50/1