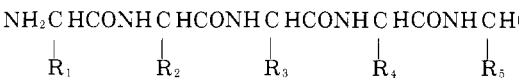
# Use of a Modified Wiswesser Notation for the Encoding of Proteins

ELIAHU HOFFMANN
The Hebrew University, Jerusalem, Israel
and
The Center of Scientific and Technological Information, National Council of Research and Development, 84 Hachashmonaim St., Tel-Aviv, Israel

Received December 20, 1968

**A modified Wiswesser Line Notation has been devised for encoding proteins. It treats the protein chain as one uniform structural entity consisting of a defined number of repeated peptide linkages. The side chains of the amino acids which make up the protein are treated as substituents of this entity and their order of appearance in the notation describes the sequence of the amino acids in the chain.**

Usually a protein (or peptide) chain can be characterized in terms of two distinct structural features, a continuous chain of peptide linkages—henceforth called the peptide chain—which consists of repeated groups of atoms of the type —CONHCH— and a series of different side groups (R)—henceforth called side chains—which are characteristic of the different amino acids which constitute the individual subunits of the chain. Thus, a small pentapeptide can be described by the following structure:



We therefore concentrated our encoding efforts on this kind of compound and disregarded for the present cyclic peptides and depsipeptides, as well as any secondary structural feature such as helical or other conformations.

As a preliminary step in the application of the notation, it seems necessary to check the compatibility of the notation rules with the structural features of the compound to be encoded and to determine whether the pertinent structural information is clearly reflected by the notation.[1]

The Wiswesser notation chooses for its primary encoding sequence that chain of connected atoms which has the most branched atoms and constitutes the longest symbol chain (in this order of priority). Thus, the peptide chain qualifies very well as a primary encoding sequence, since it will usually have both the most branched atoms and the longest chain of atoms. In addition, a protein is essentially a polymer, and as such is encoded along the polymer chain. As a polymer, it consists of a repeated sequence of atoms, the notation symbols of which conceivably can be contracted according to the pertinent WLN rules.[2] The main encoding sequence appears in notation in a greatly reduced form, relieving the complete notation of many superfluous symbols. Attention is thus drawn to the side chain notation. This tallies well with the fact that the order of appearance of the various amino acids—indicated by their side chain notation—in the protein is more important from a chemical point of view than the structure of the peptide backbone.

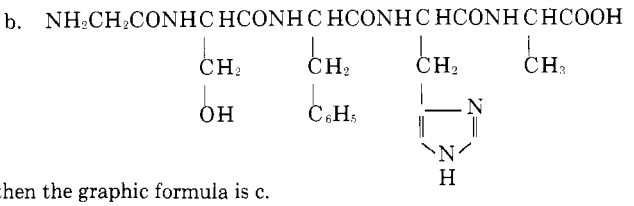
The peptide chain is expressed in notation symbols as a repeated sequence of VMY or YMV symbols, depending on the end from which the notation of the chain is started, and should therefore be cited as a multiplied group with the appropriate number—i.e., 4, as follows:

/VMY/ 4

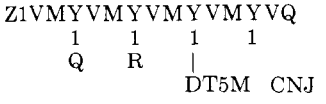
This part of the notation should be considered as one structural unit, somewhat similar to the way a heterocyclic macromolecular ring is treated. Thus, this peptide chain notation, too, has substituents—the R side chains—which are attached implicitly to the Y symbols and follow the notation just like real substituents, but without locants. Their order of citation—after the peptide chain notation—describes the sequence of appearance of the amino acids in the peptide chain. Thus, if the following pentapeptide is

a. H-Gly-Ser-Phe-Hist-Ala-OH,

or

b. 

then the graphic formula is c.



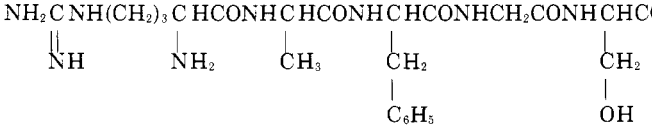
and the notation is

d. Z1/VMY/ 4 1Q 1R& 1- DT5M CNJ& 1&VQ

**Notation Analysis.** The notation starts with Z1 (in compliance with Rule 2) and is followed by the repeating unit, /VMY/, and its multiplier, 4. Then the side chain notations follow, each separated by one space. Cyclic notations are terminated by ampersands (Rule 24). The last side chain is punctuated by an ampersand, which signifies that "the symbol(s) following it is (are) attached to the same peptide unit to which the preceeding side chain notation is attached." If this side chain notation ends with a terminal symbol, the ampersand is omitted (Rule 8a). Notation symbols following such an ampersand or a terminal side chain symbol therefore represent part of the peptide chain proper, which in this specific case is the end portion of the chain. This method of citing a part of the peptide chain after a series of side chain notations will later be applied to the encoding of proline and other *N*-substituted amino acid units in a protein. These amino acids cannot be represented properly by side chain notations, and are therefore encoded in full in the peptide sequence. Glycine, however, is considered to have a hydrogen atom as a side chain, which is represented by H in a side chain notation sequence.

Citation of proteins in the literature usually starts from the amino acid end. Compliance with Rule 2 will sometimes reverse this order, as is the case with the following peptide:

a. H-Arg-Ala-Phe-Gly-Ser-OH

b. 

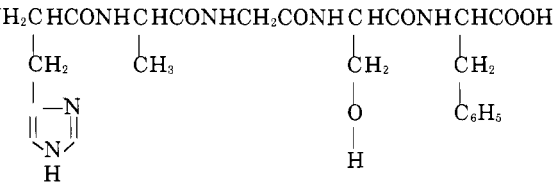
Graphic formula: c. MUYM3YVMYVMYVM1VMYVQ
Z Z 1 1 1
R Q

Notation: d. QV/YMV/ 4 1Q H 1R& 1&YZ3MYZUM

**Notation Analysis.** The notation starts with QV, being senior to MUY, and therefore the repeating unit is reversed, YMV instead of VMY. The last side chain notation (1) is again closed with an ampersand and the symbols following it are attached to the same peptide unit to which the last side chain is attached—as stated in the previous notation analysis—and constitute the end of the peptide chain proper. However, their point of attachment is not the Y symbol of the last repeating unit—as in the previous notation—but rather the V symbol of this very same unit. This, again, is due to the reversed citation of the repeating unit.

Notations may start with cyclic structures, if these are located at the ends of the peptide chain, as in the following:

a. H-His-Ala-Gly-Ser-Phe-OH

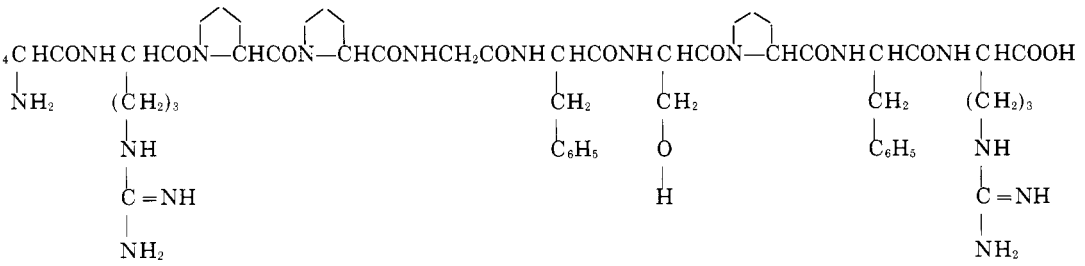
b. 

Graphic Formula: c. T5M CNJ D1YVMYVM1VMYVMYVQ
Z 1 1 1
Q R

Notation: d. T5M CNJ D1YZ/VMY/ 4 1 H 1Q VQ1R

The notation of the natural peptide Kallidin [3] illustrates the method for the incorporation of proline and other special amino acids into the peptide notation:

a. H-Lys-Arg-Pro-Pro-Gly-Phe-Ser-Pro-Phe-Arg-OH

b. 

c. Z4YVMYV- AT5NTJ BV- AT5NTJ BVM1VMYVMYV- AT5NTJ BVMYVMYVQ
Z 3 1 1 1 3
M R Q R M
YUM YUM
Z Z

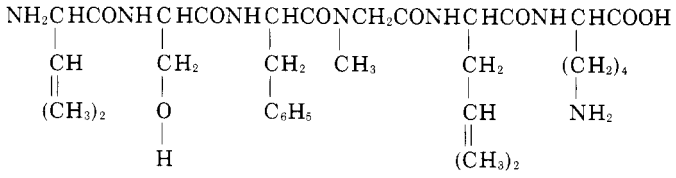
d. Z4YZ/VMY/ 6 3MYZUM&V- AT5NTJ BV- AT5NTJ B// H 1R& 1QV- AT5NTJ B// 1R& VQ3MYZUM

**Notation Analysis.** The notation starts with Z4... (senior to MUY). The first side chain 3MYZUM is closed by an ampersand, which indicates that from there on the notation represents the "chain proper"—as explained in the side chain ampersand discussions of the two previous notation analyses. Thus, the two proline units are quoted in full until B—the locant of proline at which the peptide chain starts again. Here the two slashes mean that the notation returns to its original form (/VMY/) of side chain enumeration. Thus, three side chains, H, IR&, and 1Q, follow the slashes. The terminal Q does not need an ampersand to indicate that the V symbol following it is again part of the "chain proper." From V until // the third proline unit is encoded in full, and so on to the rest of the notation. Six side chains are cited in the notation, and therefore six VMY units are cited at the beginning. This method of encoding can be used for any amino acid whose incorporation into a peptide chain does not yield a regular VMY sequence, for example, sarcosin or $\alpha$-aminoisobutyric acid, as demonstrated below:

Peptide which includes sarcosin (*N*-methyl glycin)

a. H-Val-Ser-Phe-Sarcos-Leu-Lys-OH

b.

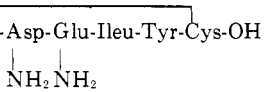


c. ZYVMYVMXVMYVM1VQ
1 1 1 1
Q R 1Y1

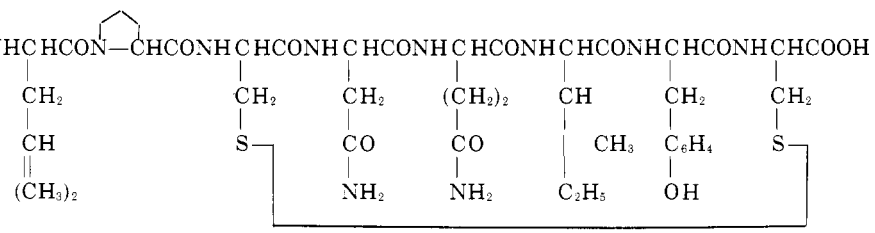
d. QV1/MVY 3 1Y&&MVX&&// 1R& Z1Q

Disulfide linkages in a protein chain, or among various protein chains, produce cyclic peptide structures. The cyclic character of these structures is chemically less important than the precise knowledge of the connection between the various sulfur atoms in the compounds. This notation, therefore, disregards the cyclic character of the resultant structure, but instead indicates the formation of the S—S bond in an unambiguous way. The notation for the natural peptide, oxytocin,[3] serves as an example.

a. H-Gly-Leu-Pro-Cys-Asp-Glu-Ileu-Tyr-Cys-OH



b.



c. Z1VMY V- AT5NTJ BVM YVMYVMYVMYVMYVMYVQ
1 1 1 2 Y 1 1
1Y1 —S V V 1 R S—
Z Z 2 Q

d. Z1/VMY/ 7 1Y&&V-AT5NTJ B// 1S- 1VZ 2VZ Y2 1R DQ& 1S-&VQ

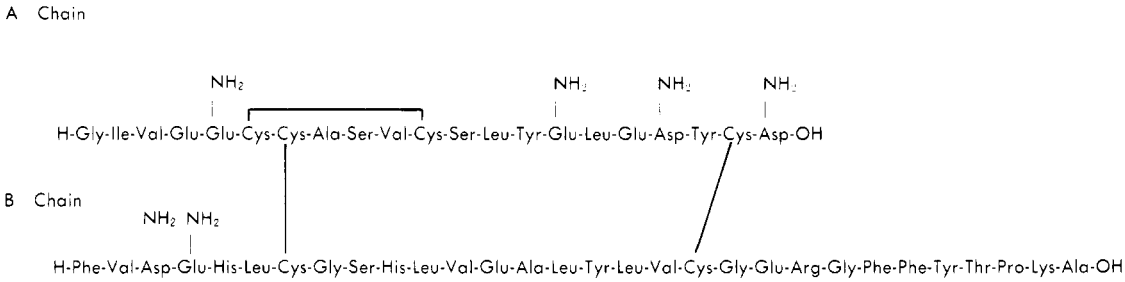
c. ZYVMYVMYVN1VMYVMYVQ
Y1 1 1 1 1 4
1 Q R 1Y1 Z

d. ZYY&&/VMY/ 4 1Q 1R&VN1&1// 1Y VQ4Z

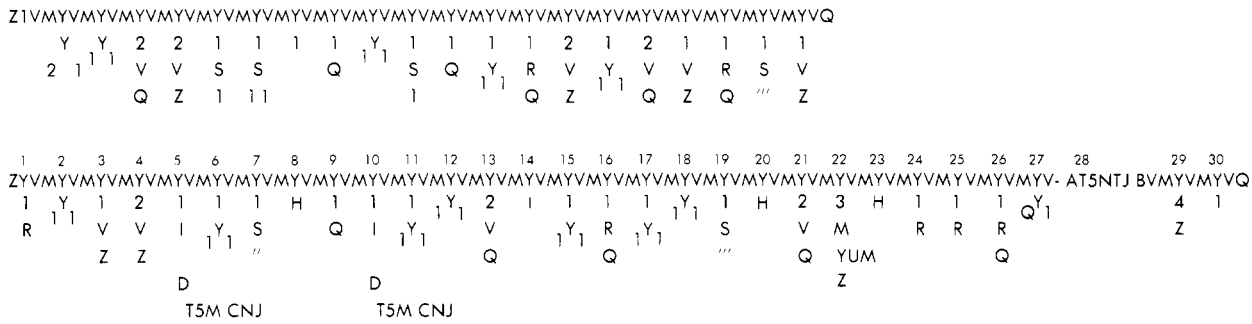
Peptide which includes $\alpha$-aminoisobutyric acid

**Notation Analysis.** The S—S linkages of the side chains in this notation are here designated by hyphens. Thus, 1S- links with 1S-. If more than one S—S linkage is present, 1S-- links with 1S--; 1S--- with 1S---, and so on.

Finally, the usefulness of this method is tested by encoding the bovine Insulin sequence[3] as follows:




Graph. Form




Notation

```
1          2 3 4  5  6  7  8 9 10 11  12 13 14      15  16 17  18 19      20 21        1          2 3  4  5
Z1/VMY/ 20 Y2 Y 2VQ 2VZ 1S- 1S-- 1 1Q Y 1S- 1Q 1Y 1R DQ& 2VZ 1Y 2VQ 1VZ 1R DQ& 1S-- VQ1VZ &R1YZ/VMY/ 29 Y 1VZ 2VZ 1- DT5M CNJ&

6  7  8 9  10           11 12 13 14 15 16      17 18 19  20 21  22      23 24  25  26      27      28      29  30
1Y 1S-- H 1Q 1- DT5M CNJ& 1Y Y 2VQ 1 1Y 1R DQ& 1Y Y 1S--- H 2VQ 3MYZUM H 1R& 1R& 1R DQ& Y&QV- AT5NTJ B// 4Z 1&VQ
```

**Notation Analysis.** The notation starts with the amino end of the A chain, Z1 senior to VQ. The B chain follows in the same direction, starting with the amino end, preceeded by an ampersand— &R1YZ/VMY/... (addition compound ampersand; Rule 15). The numbers on top of the notation do not belong to it, but are intended to identify the amino acids in the insulin formula at the top.

## SUMMARY

The method developed here uses the Wiswesser Line Notation for the encoding of peptides and proteins. It utilizes all the original symbols of the notation within the framework of the pertinent rules, and no special symbols are needed. Thus, it should be possible to use machine searching techniques for these compounds, just as for any other compound encoded according to the Wiswesser notation. The contraction of the peptide backbone has been affected by rules consistent with polymer encoding and the various amino acids and their sequence in the chain have been characterized and accentuated by the specific side chain notation.

## LITERATURE CITED

(1) Smith, E.G., "The Wiswesser Line-Formula Chemical Notation," McGraw-Hill, New York, 1968.
(2) *Ibid.*, p. 256.
(3) Meienhofer, J., *Chimia* **16**, 383 (1962).