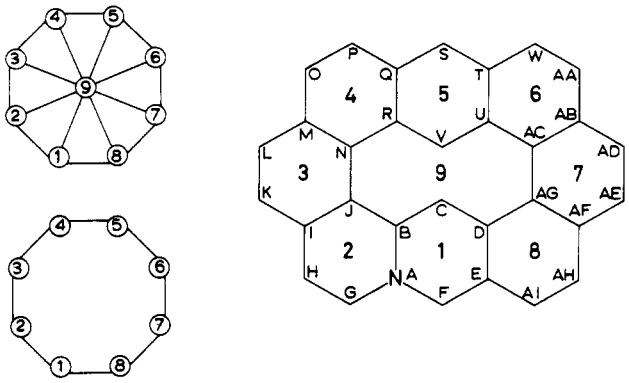
rules for encoding perifused structures are still applicable here. Since there is a closure of the ring systems, the code will contain the multi-knit part in the notation.

Example 36



ALWIN: &2<6ϕ8>A,I,M,Q,T,AB;AF2E,D,AG=NJ

Since ring 9 shares more than a single edge, it is omitted from the reduced graph (as shown). Then the spanning tree is formed as 1–2–3–4–5–6–7–8 and the elementary tree 1–2–3–4–5–6–7 and chord 7–8. Note that the notation starts with the symbols &2, since this example corresponds to ring of rings system—type II. (cf. with the ALWIN of Figure 3a).

ACKNOWLEDGMENT

The authors are indebted to Professor G. N. Ramachandran, Mr. M. Ranganathan, Dr. S. K. Sen, and Dr. K. S. N. Iyer for valuable discussions during the course of this work. The authors are also grateful to W. J. Wiswesser for several valuable suggestions. Also they thank the referees and Dr. H. Skolnik, for suggesting several improvements in the manuscript.

Two of the authors (P. V. S. and S. K.) thank the National Institutes of Health, U. S. A. (Grant No. 15964), and N.C.E.R.T. (India) for providing research fellowships.

LITERATURE CITED

(1) National Academy of Sciences and National Research Council (U. S.), "Survey of Chemical Notation Systems," Publication No. 1150, Washington, D. C., 1964.

(2) Rules for I.U.P.A.C. Notation for Organic Compounds, Longmans, London, 1961.

(3) Smith, E. G., "The Wiswesser Line Formula Chemical Notation," McGraw-Hill, New York, N. Y., 1968.

(4) Korfhage, R. R., "Logic and Algorithms," Wiley, New York, N. Y., 1966.

(5) Knuth, D. E., "Art of Computer Programming," Vol. 1, "Fundamental Algorithms," Addison-Wesley, Reading, Mass, 1968.

(6) Berge, C. E., "The Theory of Graphs and Its Application," Methuen, London, 1962.

(7) Sankar, P. V., "ALWIN—Algorithmic Wiswesser Notation for Organic Compounds," Ph.D. Thesis, Indian Institute of Science, Bangalore, 1973.

(8) Sankar, P. V., Krishnamurthy, E. V., and Krishnan, S., "Representation of Stereoisomers in ALWIN," *J. Chem. Doc.*, **14**, 141 (1974).

Representation of Stereoisomers in ALWIN†

P. V. SANKAR,** E. V. KRISHNAMURTHY,* and S. KRISHNAN

Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560012, India

Received May 22, 1973

The well-known Cahn–Ingold–Prelog method of specifying the stereoisomers is introduced within the framework of ALWIN—Algorithmic Wiswesser Notation. Given the structural diagram, the structural ALWIN is first formed; the specification symbols are then introduced at the appropriate places to describe the stereoisomers.

This paper describes the application of the Cahn, Ingold, and Prelog[1-4] method (CIP) for the representation of stereoisomers in ALWIN. It is well known that the stereoisomers have the same molecular and structural formula, but differ only in configuration or in spatial arrangement of their groups. Since the structural formula (or equivalently the connectivity table) remains identical for these stereoisomers, the line notations, such as WLN or ALWIN[5], remain identical. Therefore, if one wants to distinguish these stereoisomers, it is necessary to introduce stereochemical conventions in the form of additional punctuations in ALWIN to describe the spatial orientation or the configuration of the molecule. Stereoisomers are termed cis–trans isomers when they differ only in the position of the atoms relative to a specified plane where these atoms are parts of a rigid structure (we restrict our definition only to cis–trans isomerism about double bonds in general and to certain achiral cyclic structures whose ring skeleton approximates to a plane); these isomers are also known as geometrical isomers.

These stereoisomers which differ only in configuration are known as optical isomers because of their special property (called optical activity) of rotating the plane of polarized light. In such a case, the molecule lacks elements of symmetry, so that it is not identical with and hence not superimposable on its mirror image. Such molecules are also called chiral. Chirality of a molecule may arise because of

† Contribution No. 47, Molecular Biophysics Unit, Indian Institute of Science, Bangalore, India.

* Author to whom correspondence should be addressed.

** Presently at Tata Institute of Fundamental Research, Bombay 400005.

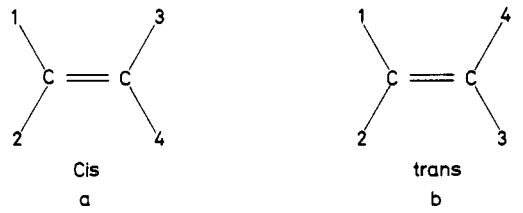


Figure 1.

the presence of a chiral center or chiral axis. We will deal with these concepts in a detailed manner later.

Also we treat the substituted monocycles and fused systems as chiral molecules (not as cis–trans isomers) in order to arrive at a unique ALWIN code.[6]

ALWIN‡ FOR CIS–TRANS ISOMERS

Acyclic Structures

ACS 1

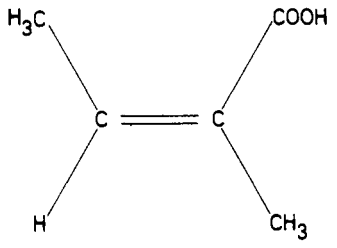
Whenever the acyclic structure contains a double bond, the following steps are used for forming the ALWIN code.

(i) Form the structural ALWIN (for the sake of clarity we refer to ALWIN formed from structural formula using the rules of ref 5 as structural ALWIN).

(ii) Identify the first neighbors of double bond atoms; among the pairs of such atoms at either end choose the reference atoms (see Blackwood, *et al.*[7]) satisfying the sequence rule of CIP (see Appendix).

(iii) Depending on whether these atoms are on the same side (*e.g.*, atoms 1 and 3 in Figure 1a) or on the opposite side (*e.g.*, atoms 1 and 3 in Figure 1b) cite the symbols /Z/ and /E/, respectively, in the notation immediately before U (double bond under consideration). We have enclosed Z and E between two slashes to differentiate these from Z = NH_2 and E = Br used in ALWIN.[5] Thus we can supress these symbols for alphabetical ordering. Later, these stereoisomers can again be ordered with /Z/ having a higher precedence than /E/. This would give us a list of all the stereoisomers together as one group in an alphabetical listing, without being separated.

Example 1



ALWIN : QVYA/Z/U@A

The structural ALWIN for this example is QVYAU@ A. The Sequence Rule prefers the groups COOH and CH_3 at either end; they are cis with respect to each other.

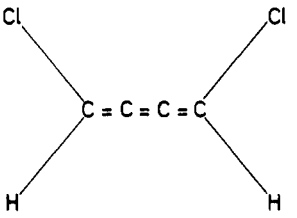
ACS 2

If there are two or more ways of writing the notation, choose that notation which cites /Z/ first.

ACS 3

In the case of cumulenes (odd number of cumulative double bonds), locate the terminal carbon atoms of such a sequence of carbon atoms joined by successive double bonds. Use rule ACS 1 to assign the configuration of the isomer, choosing the reference atoms (groups) from among the immediate neighbors of these terminal carbon atoms. As a convention, cite the symbol /Z/ or /E/ immediately before the symbol U corresponding to the last double bond to be cited in the notation.

Example 2



ALWIN : G@UCUC/Z/U@G

Here, both the terminal atoms have the neighbors Cl and H; among them the chlorine atoms are chosen as the reference atoms.

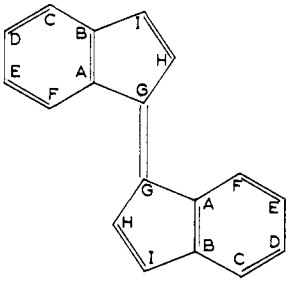
Cyclic Structures

CS 1

The configuration about the double bond in a cyclic structure is decided the same way as in ACS 1. However, here the symbol /Z/ or /E/ is cited between the symbol U corresponding to the bond under consideration and its corresponding locant (in the ring atom zone of the IWF (Information Word Format[5]) if this double bond joins two ring atoms). In the case of unsaturated ring systems, the double bonds are not cited in the notation; however, if a particular double bond in an unsaturated system gives rise to stereoisomers, then the symbol U must be cited in the notation.

If a substituent is connected to a ring atom by a double bond, then cite the symbol /Z/ or /E/ after the ring locant in the substituent zone.

Example 3



ALWIN : L65J&G/E/U-&GL65J

In this example, the ring atoms with locant A in either ring are chosen as the reference atoms by the Sequence Rule (Appendix).

CS 2

On forming the code for cyclic structures, if after the application of the precedence rules there are equivalent ways of labeling the ring atoms, then choose that labeling which gives the highest locant to the double bond with a Z conformation.

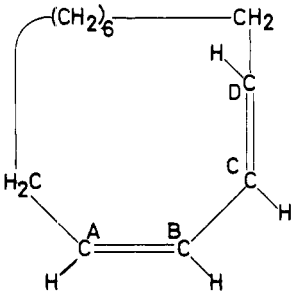
In example 4, the atoms joined by the double bond which gives rise to a cis conformation are assigned the highest locants.

CS 3

In the case of cyclic structures like 1,*n*-dialkylidenecycloalkanes, in which the cycloalkanes are symmetrically substituted and have $2n - 2$ members where $n = 3, 4, 5 \ldots$,

‡ The rules will be prefixed by ACS or CS according as they refer to acyclic or cyclic structures, respectively. In this paper we use parentheses-free ALWIN for acyclic structures.[5]

Example 4



ALWIN: L-12-T/Z/U&C/E/UJ

the cis–trans isomerism can be defined in much the same way as for the simple double bond systems.

The ALWIN notation for such a system is formulated using the following steps.

(i) Write the structural ALWIN.

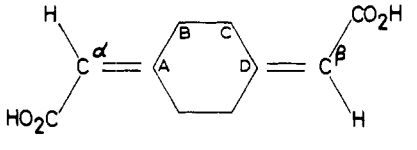
(ii) Identify the terminal atoms (by terminal atoms, we refer to those atoms which are joined by double bond to the cycloalkane).

(iii) Consider the connected atoms of these terminal atoms and use rule ACS 1 to assign the configuration.

(iv) In the notation, cite the symbol* % between the symbol U and the locant of the terminal atom which is connected to the ring atom with a higher locant.

(v) Cite the appropriate symbol /Z/ or /E/ between the symbol U and the locant of the other terminal atom.

Example 5



ALWIN: L6TJ&A%U@VQ&D/E/U@VQ

The terminal atoms in this example are marked α and β, respectively. By applying ACS 1, we find that the two COOH groups are in the trans conformation.

ALWIN FOR MOLECULES WITH A CHIRAL CENTER

An asymmetric atom is termed a chiral center if the ligands (generally four in number) connected to it are all different. This is denoted by the set of symbols Cabcd. The ligands a, b, c, and d can be arranged in a definite order of decreasing precedence as a > b > c > d using the Sequence Rule (see Appendix). Now looking at the 3D model from the least ligand d, if the sequence a, b, and c traces a clockwise course (Figure 2a), then the particular configuration is assigned the label R; if the sequence a, b, and c traces an anticlockwise course (Figure 2b), then the corresponding configuration is assigned the label S.

Acyclic Structures

ACS 4

If an acyclic structure has an asymmetric atom, then form the ALWIN code using the following steps.

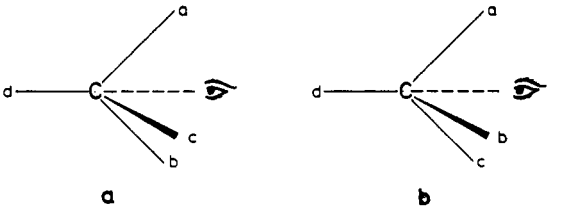
(i) Write the structural ALWIN.

(ii) Order the ligands connected to the asymmetric atom as a > b > c > d, using the Sequence Rule.

(iii) Assign the configurational label R or S depending on whether the ligand sequence a, b, and c forms a clockwise or anticlockwise course looking from the least preferred ligand d.

(iv) In the notation, introduce the appropriate symbols /R/ or /S/ just before the notation symbol of the asymmetric atom under consideration.

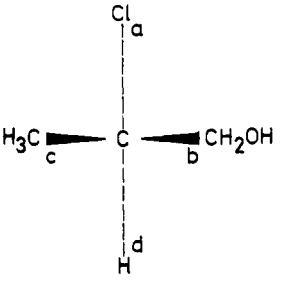
* The symbol % acts as a tag for the atom with respect to the conformation at the other terminal atom.



a b

Figure 2.

Example 6



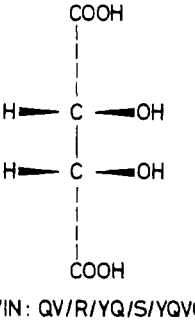
ALWIN: Q1/R/YAG

We use dotted lines for indicating bonds going inside and thickened lines for bonds coming out of the plane of the paper. The groups a, b, c, and d are respectively marked in the structural diagram; this isomer is in the *R* configuration.

ACS 5

If there are two or more ways of writing the ALWIN notation for any structure, then choose the notation which cites /R/ first.

Example 7



ALWIN: QV/R/YQ/S/YQVQ

In this example, the structural AAALWIN remains the same if we write the code starting from either COOH group; however, we choose the above ALWIN code by rule ACS 5.

Cyclic Structures

CS 4

Assign the configuration of the asymmetric centers in the cyclic structures using rule ACS 4. Introduce the stereochemical symbols in the structural ALWIN as follows.

If the asymmetric atom is a ring atom, then cite the symbol /R/ or /S/ immediately before the notation symbol of this atom in the ring atom zone. If this ring atom happens to be a carbon atom, then cite the appropriate symbol X or Y in the notation after the symbol /R/ or /S/ (usually such

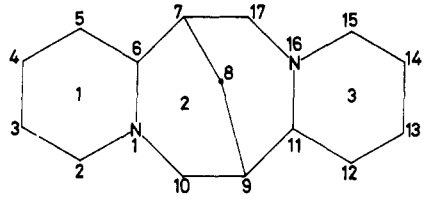
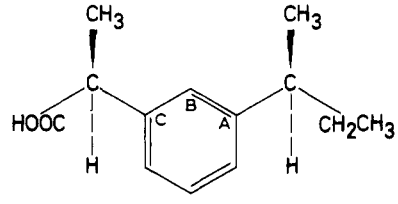


Figure 3.

carbon atoms are not cited in the notation). In the case of spiro-carbon atoms which are asymmetric, the symbol X is cited in the notation, even though this information might perhaps be implicit in the tessellation code.[5]

Example 8



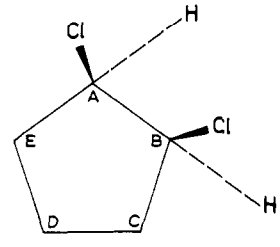
ALWIN : RϕA/R/YA1AϕC/R/YAVQ

In this example, the two asymmetric atoms connected in the benzene ring at locants A and C are respectively in *R* form.

CS 5

In case the ring atoms can be labeled in two or more ways, after the application of the precedence rules, assign the locants to the ring atoms such that the asymmetric atoms with *R* configuration are given the highest locants.

Example 9



ALWIN : L5TJϕA/R/GϕB/S/G

In example 9, of the two asymmetric centers, one is in the *R* configuration, while the other is in the *S* configuration. By rule CS 5, the former is given the locant A.

The planar projection of example 10 is shown in Figure 3. The given structure is a bridged ring system with the perifused ring as the main ring system. Either of the terminal rings (the terminal nodes 1 and 3 in the reduced graph) can be assigned the locants first since they give rise to the same ALWIN notation.

The locant assignment is shown in Figure 4.

Pseudoasymmetry. An atom is termed pseudoasymmetric when it is tetrahedrally bound to two enantiomeric groups, and the other two groups are different from each other; *i.e.*, they are not even enantiomeric to each other.

S 1

The Sequence Rule is applied to order the ligands connected to a pseudoasymmetric center, and rule ACS 4 is applied to assign its configuration. This is denoted in the notation by the symbols /R# or /S#.

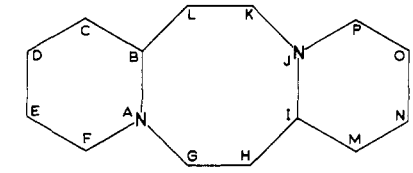
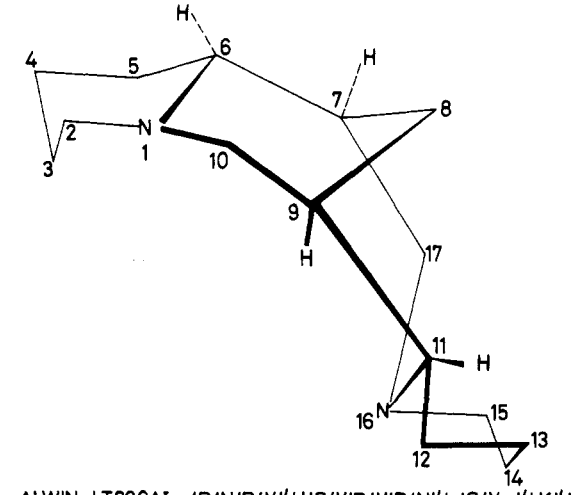


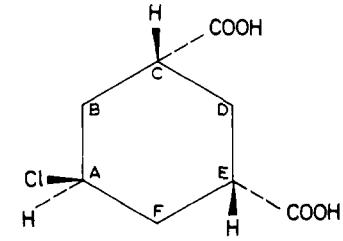
Figure 4.

Example 10



ALWIN : LT686AI = /R/N/R/YϕH/S/Y/R/Y/R/NϕL/S/Y : ϕH1ϕLJ

Example 11



ALWIN : L6TJϕA/R # GϕC/R/VQϕE/S/VQ

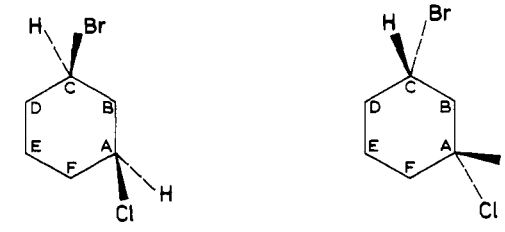
In example 11, the ring atom with chlorine as its substituent is assigned the locant A. Among the ring atoms to which the COOH groups are connected, the one with the *R* configuration is assigned the highest locant, say C (by rule CS 5). The ring atom at locant A is pseudoasymmetric and has the configuration /R# by rule S 1.

Relative Configuration

S 2

In compounds containing chiral centers of which only the relative but not the absolute configurations are known, the relative configurations are denoted in the notation by the symbols /R% and /S%.

Example 12



ALWIN : L6TJϕA/R %GϕC/S%E

ALWIN FOR MOLECULES WITH A CHIRAL AXIS

Certain allenes, biaryl compounds, and spiranes exhibit optical isomerism, even though they do not contain chiral centers; this is due to the presence of a chiral axis in the structure as a whole. The concept of the chiral axis is less stringent, as far as the difference between the connected ligands are concerned, than the chiral centers. Since these concepts have been dealt with in detail by Cahn, Ingold, and Prelog, these are not discussed here.

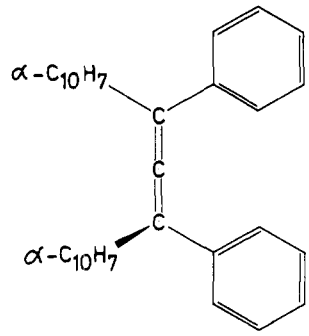
Allenes

ACS 6

In the case of allenes containing an even number of cumulative double bonds, use the CIP convention to assign the configuration.

In the notation, cite the appropriate symbol /R/ or /S/ immediately before the notation symbol corresponding to the last atom (of such a series of carbon atoms joined by cumulative double bonds) cited in the notation.

Example 13



ALWIN: R␣AY9AUCU/R/Y9AR

Cyclic Structures

CS 6

In ring structures obtained by replacing one double bond of an allene by a ring, the ALWIN is formed as follows.

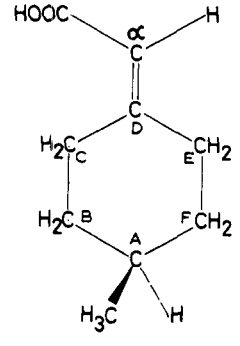
(i) Write the structural ALWIN.

(ii) Assign the configuration using the CIP convention.

(iii) In the structural ALWIN, introduce the symbol % before the notation symbol of that terminal atom (here, we mean by terminal atoms those atoms whose substituent groups were ordered by the Sequence Rule for assigning the configuration) which is connected directly or through a sequence of atoms to the ring atom with the higher locant (the terminal atom itslf can be a ring atom); now cite the appropriate symbol /R/ or /S/ before the notation symbol of the other terminal atom.

(iv) If the terminal atom, apart from being a ring atom, also happens to be a carbon atom, then cite the appropriate symbol X or Y in the notation.

Example 14



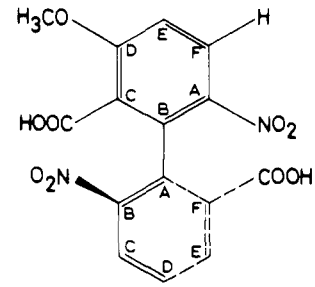
ALWIN: L6T%YJ␣AA␣DU/R/YVQ

In example 14, one of the terminal atoms is the ring atom with the locant A; the other terminal atom is labeled α; this structure is in the *R* form.

CS 7

In the case of biaryls, use the CIP convention to assign the configuration. In the notation, cite the appropriate symbol /R/ or /S/ immediately after that ring locant of the main ring to which the substituent ring is connected.

Example 15



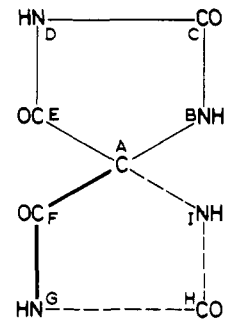
ALWIN: R␣ANW␣B -(/S/R␣BNW␣FVQ)␣CVQ␣DOA

In the notation the symbol /S/ is cited immediately after the locant B of the main (benzene) ring.

CS 8

In the case of spiranes, cite the symbol /S/ or /R/ immediately after the locant of the spiro atom in the ring atom zone. Also, cite the symbol X immediately following it.

Example 16



ALWIN: LT55/S/XMVMVVMVMJ

Contracted as LT55/S/X$<MVMV>J

ACKNOWLEDGMENT

Two of the authors (P. V. S. and S. K.) thank the National Institutes of Health, U. S. A. (Grant No. 15964) and N.C.E.R.T., India, for providing research fellowships. The authors are indebted to Professor G. N. Ramachandran for many valuable discussions.

APPENDIX

The Sequence Rule. The ligands associated with an element of chirality are ordered for comparing them at each step in bond-by-bond exploration, from the element, along the successive bonds of each ligand; however, if the ligands branch, then the exploration is carried first along the branch paths providing the highest precedence to their respective ligands, and then the exploration is continued to order all the groups using the following standard subrules each to exhaustion, in turn, namely:

(0) Nearer end of axis or side of plane precedes farther.

(1) Higher atomic number precedes lower.

(2) Higher atomic mass number precedes lower.

(3) Seqcis precedes seqtrans.

(4) Like pair R,R or S,S precedes unlike pair R,S or S,R; M,M or P,P precedes M,P or P,M; R,M or S,P precedes R,P or S,M; M,R or P,S precedes M,S or P,R; also r precedes s.

(5) R precedes S; M precedes P. Note: The above notations have the following correspondence with our notation: R ≡ /R/; S ≡ /S/; r ≡ /R#; s ≡ /S#. Also M and P denote left-handed and right-handed helices, respectively.

LITERATURE CITED

(1) Cahn, R. S., and Ingold, C. K., "Specification of Configuration about Quadricovalent Asymmetric Atoms," *J. Chem. Soc.*, 612 (1951).

(2) Cahn, R. S., Ingold, C. K., and Prelog, V., "Specification of Configuration in Organic Chemistry," *Experientia*, **12**, 81 (1956).

(3) Cahn, R. S., Ingold, C. K., and Prelog, V., "Specification of Molecular Chirality," *Angew. Chem. Int. Ed. Engl.*, **5**, 385 (1966).

(4) Cahn, R. S., "An Introduction to the Sequence Rule," *J. Chem. Educ.*, **41**, 116 (1964).

(5) Krishnamurthy, E. V., Sankar, P. V., and Krishnan, S., "ALWIN—Algorithmic Wiswesser Notation System for Organic Compounds," *J. Chem. Doc.*, **14**, 130 (1974).

(6) IUPAC "Tentative Rules for the Nomenclature of Organic Chemistry, Section E, Fundamental Stereochemistry," *J. Org. Chem.*, **35**, 2849 (1970).

(7) Blackwood, J. E., Gladys, C. L., Petrarca, A. E., Powell, W. H., and Rush, J. E., "Unique and Unambiguous Specification of Stereoisomerism about Double Bond in Nomenclature and Other Notation Systems," *J. Chem., Doc.*, **8**, 30 (1968).

Huffman Binary Coding of WLN Symbols for File-Compression†

K. SUBRAMANIAN,‡ S. KRISHNAN, and E. V. KRISHNAMURTHY*

Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560012, India

Received December 3, 1973

Construction of Huffman binary codes for WLN symbols is described for the compression of a WLN file. Here, a parenthesized representation of the tree structure is used for computer encoding.

In this paper, we describe the construction of Huffman binary codes[1] for WLN symbols, to facilitate file-compression in storing WLN for chemical compounds. This construction has been made feasible by the data given in ref 2 which gives the statistics of occurrence of WLN symbols.

It is well known that the Huffman procedure yields an optimum set of uniquely decipherable and instantaneously decodable code words (meaning thereby that no other set has a smaller number of symbols per message); hence the code constructed is of value in economizing the chemical information storage.

HUFFMAN CODE

Given a set of N source symbols $S = \{s_1, s_2, \ldots, s_N\}$ and their probability of occurrences $P = \{p_1, p_2, \ldots, p_N\}$, the Huffman procedure constructs a set of instantaneously decodable (optimum length) code words using the code alphabet set $C = \{c_1, c_2, \ldots, c_M\}$ by the following steps.

Step 1. The N symbols are arranged in the order of their decreasing probabilities.

Step 2. Let k be the integer satisfying the two requirements

$$\begin{aligned} &\text{(i)} \quad \text{For } M = 2,\ k = 2 \\ &\text{(ii)} \quad \text{For } 3 \leq k < M \end{aligned} \tag{1}$$

$$(N - k)/(M - 1) = \text{positive integer} \tag{2}$$

Here we add β $(=(N - k) \text{ MOD } (M - 1))$ dummy symbols (to the given symbol set), each with a zero probability of occurrence.

Group together k least probable symbols and compute the total probability of this subset.

Step 3. Construct an auxiliary ensemble of symbols from the original ensemble, by regarding the subset of k symbols formed in step 2 as a single symbol, with probability equal to the probability of the whole subset. Rearrange the symbols of the auxiliary ensemble in the order of their decreasing probabilities.

Step 4. Again form a subset of the k least possible symbols of the auxiliary ensemble and compute its total probability.

Step 5. Construct a second auxiliary ensemble from the first auxiliary ensemble, by regarding the subset of k symbols formed in step 4 as a single symbol, with probability equal to the total probability of the whole subset. Rearrange the symbols of this second auxiliary ensemble in the order of decreasing probabilities.

Step 6. Form successive auxiliary ensembles by repeating step 4 and step 5 until the total number of symbols in the auxiliary ensemble is equal to the number of code alphabets M.

Step 7. The preceding steps, when formally carried out, yield a tree for which the specified messages are the terminal nodes. Code words can be constructed by assigning different symbols from the prescribed alphabet to the branches stemming from each intermediate node.

PARENTHESIZED FORM OF HUFFMAN PROCEDURE

For computer implementation, we use a parenthesized representation (PR) of the tree structure encountered in Huffman coding;[1] this facilitates easy computer parsing. This is illustrated by an example below. Let

† Contribution No. 48, Molecular Biophysics Unit, Indian Institute of Science, Bangalore, India.

‡ School of Automation, Indian Institute of Science, Bangalore-12, India.

* Author to whom correspondence should be addressed.