SUMMARY

The rules of precedence of the Wiswesser Line Formula Notation for polycyclic structures are of such a nature that the canonical notation for complex structures is sometimes obtained only with time-consuming effort. A computer program has been written which accepts as input a set of nonconsecutive locant pairs which unambiguously describes a cyclic structure. The program operates on this information and produces a detailed analysis of the network according to the rules of the notation. The correct canonical notation is also prepared by the program. As a by-product of this analysis, the program punches on cards the nonconsecutive locant pair sets for each one of the paths tried. These data are used to accumulate an index of the structures analyzed and thus prevent unnecessary duplication of effort.

LITERATURE CITED

(1) Bowman, C. M., Landee, F. A., and Reslock, M. H., "A Chemically Oriented Information Storage and Retrieval System. I. Storage and Verification of Structural Information," J. CHEM. DOC. 7, 43 (1967).

(2) Landee, F. A., "Computer Programs for Handling Chemical Structures Expressed in the Wiswesser Notation," Presented before the Division of Chemical Literature, 147th National Meeting of the American Chemical Society, Philadelphia, Pa., April 8, 1964.

(3) Patterson, A. M., Capell, L. T., and Walker, D. F., "The Ring Index," 2nd ed., American Chemical Society, 1960; Suppl. **I**, 1963; Suppl. **II**, 1964; Suppl. **III**, 1965.

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

Structure Display*

ERNEST HYDE and LUCILLE THOMSON
Imperial Chemical Industries, Ltd., P.O. Box 25,
Alderley Park, Macclesfield, Cheshire, England

Received May 21, 1968

Structure display as the end point of searching is essential if mechanization is to be acceptable to the organic chemist. The standard of display achieved must be as compatible as possible with current practice. Any system proposed for mechanization is incomplete without this facility. This paper gives details of a program suitable for the regeneration of a structure from a connectivity matrix derived from the Wiswesser notation.

An investigation has been carried out to establish the suitability of the Wiswesser linear notation for computer systems.

A notation technique is based on describing the structural features of a chemical compound by a series of symbols whose order is governed by a set of precise rules. The resulting expression gives a unique representation of a compound in a form well suited to building up a compound registry. In its original form, the notation is not entirely suitable as a method of representing structural data for computer manipulation. However, the notation can be converted by computer into a connection table, a form which does give a precise chemically descriptive record suitable for all computer applications.

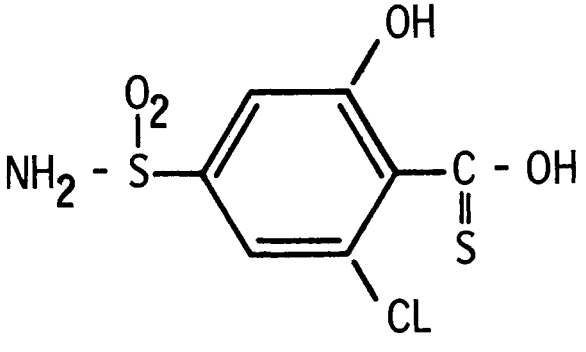
The connection table developed from the notation is written in a linear form and consists of three parts:

Chemical Unit Section. A linear string of symbols each representing an atom and its associated bond. The chemical units are assumed to be linked together in a linear string. This assumption is modified by the data contained in the second part of the table.

Connection Transfers. A set of numbers written in pairs which indicate: a) the unit position where a break in connectivity has occurred and b) the position of the branch from which the connectivity pattern is to be resumed.

Ring Block. This is the last section of the connection table and indicates those units which form a ring system. The data are expressed as a set of numbers that represent the pairs of atoms forming each ring closure.

For example,



* Presented before the Division of Chemical Literature, Symposium on Notation Systems, 155th Meeting, ACS, San Francisco, Calif., April 4, 1968.

The Wiswesser notation for this molecule is

ZSWR CQ EG DYQUS

The connection table form of the notation is

1	2	3	4	5	6	7	8	9	10	11	12	13	14
Z	S	W	T	D	T	T	T	D	Q	G	U	Q	S

1) Chemical Units
2) Connection Transfers 03.02, 09.06, 10.08, 11.07, 13.12
3) Ring Block 04.09

The connection table has two distinctive features which are important for effective computer manipulation.[1]

Compactness of Symbol Definition. The chemical units contained in the connection table represent both an atom and its associated bond. Compacting an atom and bond into a single unit results in a more highly defined node for computer recognition. Also, the derivation of the connectivity network for an entire molecule is made more simple, since the step of establishing atom-bond connections has been eliminated.

Precise Expression of Connectivity. The connection transfer data indicates where breaks in connectivity have occurred in the assumed linear string of chemical units. Rather than listing all unit pair connections within a molecule this approach assumes that the units are linked together in a continuous string and modifies this assumption when it is not correct. The resulting connectivity statement is greatly condensed since only those unit positions which are terminal or which branch are designated. This format provides rapid and easy access to the total connectivity network of a molecule, providing the exact "stop" and "pick-up" points for the linear string of chemical units.

To date, two applications of the connection table form of the notation have been demonstrated:

The Generation of Open-Ended Fragment Files. A previous paper[2] has described the results of a study which concerned the generation of fragments by computer, using as input the connection table representation of a molecule. This technique results in an open-ended coding system and is flexible in the sense of permitting a user to define a particular fragmentation code.

The Generation of Structure Display. Several investigational programs have been written to demonstrate the feasibility of generating an acceptable 2-dimensional display for output on a line-printer. The results of the study have established that it is possible to generate a structural diagram by computer using the connection table form of the notation as input. This paper will describe some of the computer techniques developed during the investigation.

The display program has been written in COBOL for an IBM 360 computer system. A full description of this program is difficult because of its complexity and size. (The program contains over 800 COBOL statements.) For this reason, a simple account of the important display principles will be presented, using an example to illustrate these points.

COMPUTER STEPS IN STRUCTURE DISPLAY

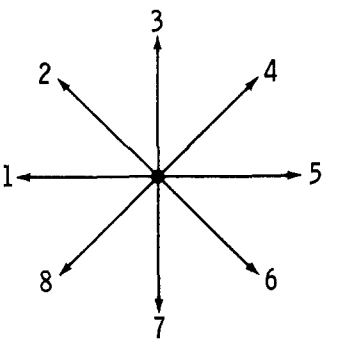
There are four steps involved in the generation of a structural diagram from a connection table

1) The program must derive the connectivity network for the entire molecule using the connection transfer data.
2) The next step involves the ring block section of the table. The program marks the presence of ring units (if any) and notes the positions of ring-closure units with a special indicator.
3) The chemical units are translated into their normal atomic representation as the final step is executed.
4) This last step, the most complex part of the computer program, is concerned with the plotting of each atom and bond according to a free plotting routine. This routine actually computes the (X,Y) coordinates required to position each atom and its associated bonds correctly. Positioning linking atomic groups is a simple case of linear plotting. Directional changes become necessary when a branching unit is approached. The programs derive the appropriate angles or "new directions" of tracking to plot atoms and bonds of side branches. Ring atoms require special consideration as well. The routine must compute the (X,Y) coordinates for ring atoms and bonds to give a closed ring diagram. The complexities in plotting arise in attempting to overcome any overwriting errors. The program must recognize a situation when an overwriting conflict is likely to occur and then have the ability to modify its tracking route to avoid the error.

FREE PLOTTING PRINCIPLES

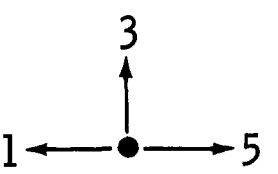
The (X,Y) coordinates of points required to print the structural diagram are derived and plotted in a "grid-area" field defined in the memory of the computer. When the plotting has been completed, a line-by-line printout of this grid-area commences to provide the 2-dimensional structure as output.

The free plotting routine allows for eight possible tracking directions. They are:

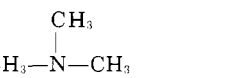


The program can select any combination of these routes. At a branching unit, a combination of these routes is selected to plot side branches. For example,

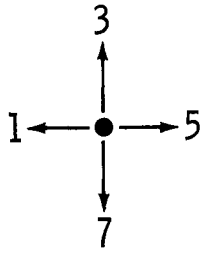
A 3-branch structure would require directions



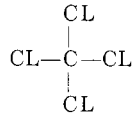
drawn as



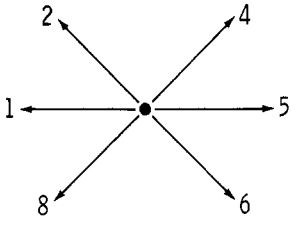
Similarly a 4-branch structure would require



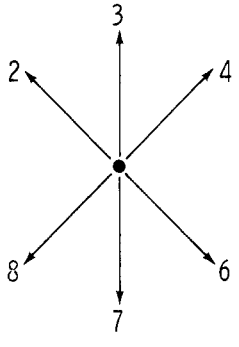
drawn as



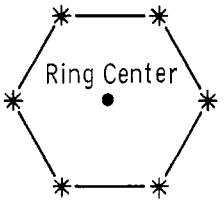
Rings are plotted by selecting a combination of these same 8 directions. The ring center is treated as branching point and each ring atom is plotted along 1 of the 8 available directions. For example, a 6-membered ring would require the directions,



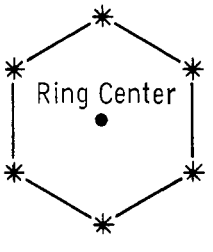
or



which is drawn as

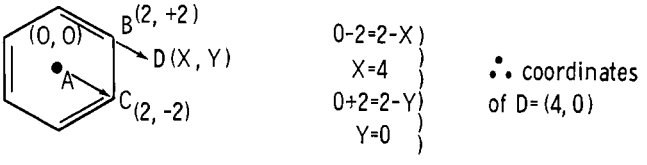


or



By a selection of the available directions, it is possible to draw various sizes of rings in either a horizontal or vertical form (see Appendix I). Fused ring systems are drawn by a free plotting technique which is based on drawing the first ring, then computing the ring center for the second, drawing the second ring, the computing the ring center of the following ring, etc. The ring center is computed knowing the coordinates of the previous ring center and coordinates of the ring junction atoms. For example,

Coordinates of D are computed such that line AC is parallel to BD

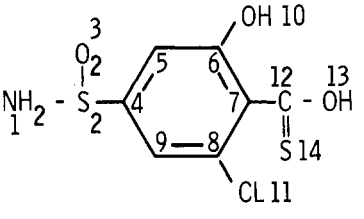


A = ring center of ring no. 1
B & C = ring atoms at the junction
D = ring center of ring no. 2

Examples of fused ring systems generated in the programs are given in Appendix II.

ILLUSTRATION OF STEPS INVOLVED IN GENERATED DISPLAY

Consider the previous example,



whose connection table representation is:

1	2	3	4	5	6	7	8	9	10	11	12	13	14
Z	S	W	T	D	T	T	T	D	Q	G	U	Q	S

Chemical Units
Connection Transfers 03.02, 09.06, 10.08, 11.07, 13.12
Ring Block 04.09 0000—i.e., units in positions 4–9 are ring atoms

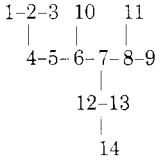
Step 1. Derive Connectivity Network. The connection transfer data gives the "stop" and "pick-up" positions in the linear string of chemical units.

The connection transfers are written in pairs: the first number shows the break in connectivity; the second gives the pick-up point.

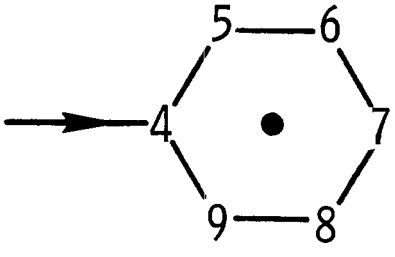
In the example under consideration, the data can be represented as:

Connection Transfers		Unit Connections
Stop	Pick-up	
03	02	1-2-3 (stop pick-up at 2)
09	06	2-4-5-6-7-8-9 (stop pick-up at 6)
10	08	6-10 (stop pick-up at 8)
11	07	8-11 (stop pick-up at 7)
13	12	7-12-13 (stop pick-up at 12)
		12-14

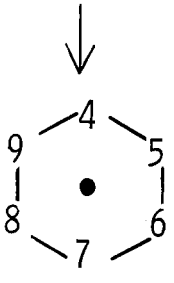
The connectivity network expressed as unit positions is:



Step 2. Mark Ring Units. The ring block indicates the presence of a six-membered ring represented by units 4–9 with units 4 and 9 forming the ring closure. Depending on the direction of approach the ring will be plotted in a horizontal form as in



or in a vertical form as in



Step 3. Translate Units to Atom Representation. Translation of chemical unit symbols to conventional atom representation requires a look-up in one of two tables.

Table 1. Units for Acyclic Structures

Chemical Unit	Atom Representation	Associated Bonding
D*	CH	—CH= or =CH—
E	BR	—BR
G	CL	—CL
J*	N	—N= or =N—
K	N	—N<
L	CH_2	—CH_2—
M	NH	—NH—
Q	OH	—OH
R	Phenyl	Phenyl
U*	C	=C— or —C— or —C=
V	CO	—CO—
W	O_2	—O_2
X	C	—C<
Y	C	—C—
Z	NH_2	—NH_2
#	N	—N≡ or ≡N—

Table 1 shows the 16 unit definitions required for the translation of the acyclic parts of a molecule. This is a relatively small number, and the reason is that there are several units which do not require an entry in this table, since they are represented by the conventional atomic symbol. These are:

Unit and Atom	Associated Bonds
B	—B=
C	—C≡ or ≡C—
F	—F
H	—H
I	—I
N	—N—
O	—O—
P	—P— or —P—
S	—S—, —S—, —S—

The translation of units to atoms is a simple step, requiring a look-up in a table of only 16 entries. The insertion of bonds between atoms is also a straightforward step owing to bond order definition within the unit symbol. Those units marked by an asterisk in the table (D, J, and U) are the only three symbols which require a double bond. When these units are connected together, alternating double bonds are inserted, always commencing with a double bond.

Units J and D are either linking or terminal and appear as:

UNIT	TERMINAL	LINKING
D	$=CH_2$ or $CH_2=$	=CH— or —CH=
J	=NH or NH=	=N— or —N=

Units D and J can only take double bond connections with each other and U. Unit U is always branching and can appear as

X=U— or —U— or —U=X

where X is {D, J, or U and terminal S or O} as in —C(=S)—OH or —C(=O)—OH

The handling of triple bonds is simple as well. There are only 2 units which require a triple bond. These are:

C, which is —C≡ or ≡C— and # which is —N≡ or ≡N—

Examples of compounds for which double and triple

bonds have been generated and printed are given in Appendix III.

The acyclic sections of the example would require the following conversions:

Unit Position	Unit	Symbols for Display
1	Z	NH_2
3	W	O_2
10	Q	OH
11	G	CL
12	U	C
13	Q	OH

Table 2. Units for Cyclic Structures

Chemical unit	Atom representation	Associated bonding
D	•	- • = or = • -
J	N	- N = or = N -
L	•	- • -
M	NH	- NH -
T	•	- • = or = • -
U	•	- • -
X	•	- • -
Y	•	- • -

• = Carbon

Those units which represent a carbon atom in a ring are displayed as a heavy dot (•). In all, only 3 symbol conversions are required:

D, L, T, U, X, Y → •
J → —N=
M → —NH—

There are several unit symbols which do not require translation since they are represented by the atomic symbol. They are:

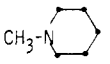
N
O
P
S

As in acyclic chains, the bonds required between ring units are obtained from the unit definitions. Those ring units which require single bonds are tagged and alternating single and double bonds are then inserted between the remaining ring units, always starting with a double bond.

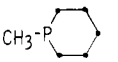
Those ring units which always require single bonding are:

L (-CH_2-) as in 

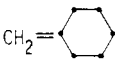
M (-NH-) as in 

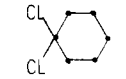
N (-N-) as in 

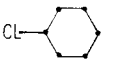
O (-O-) as in 

P (-P-) as in 

S (-S-) as in 

U (-•-) as in 

X (-•-) as in 

Y (-•-) as in 

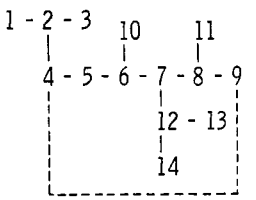
The display symbols and bonds required for the ring in the example are given below:

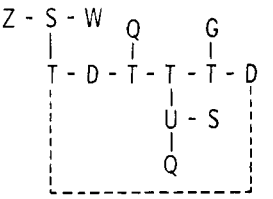
Unit position	Unit	Display symbol	Bonding
4	T	•	
5	D	•	
6	T	•	
7	T	•	
8	T	•	
9	D	•	

Step 4. Plotting Atoms and Bonds. The atom-bond pattern is constructed using that data generated in the previous steps. The (X,Y) coordinates are derived for each atom and bond, and these symbols are plotted in a defined grid-area.

Plotting commences with unit 1 and a left-to-right horizontal motion is assumed. This direction will be altered when plotting the start of a side branch and in plotting rings. Having completed the alteration, the horizontal left-to-right direction is always resumed.

The order in which units are plotted is governed by the connectivity network derived in Step 1.



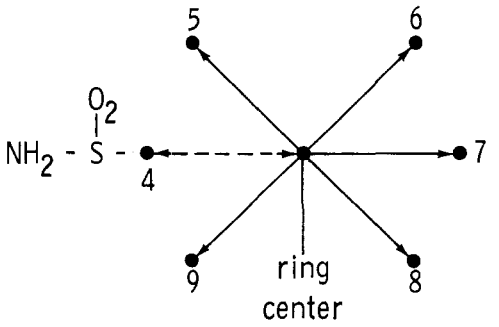


Making use of the unit conversions obtained in step 3 plotting units 1, 2, and 3 gives

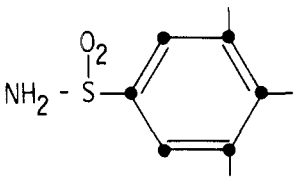
Left → Right

NH_2—S(=O_2)→

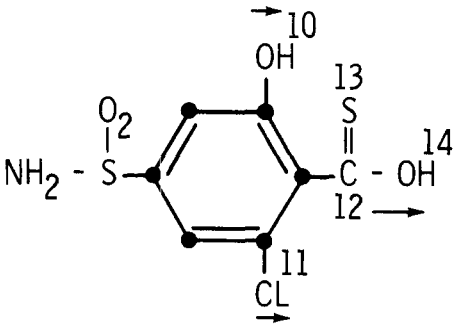
An examination of the ring block data in step 2 resulted in the tagging of units 4 and 9 which form the ring closure. The direction of approach from the S atom to ring unit 4 is horizontal. An imaginary line is drawn from the sulfur atom, through the T unit (unit 4) to an imaginary point called a ring center. This point is treated as a branch, and six of the available eight directions are selected to plot units 4 to 9.



Making use of the ring bond pattern developed in step 3, the structure can be drawn as



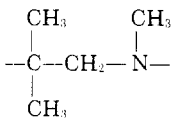
Having established the direction of approach from the ring, the ring substituents are plotted, assuming a left-to-right horizontal motion.



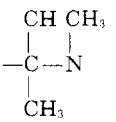
This example illustrates a relatively simple display problem. There are additional considerations in structure display which have not been covered by this example. One of the most difficult problems in generating structure display is to recognize and overcome any overwriting situation.

THE OVERWRITING PROBLEM

Consider the simple case of plotting branched structures as in,

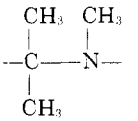


An overwriting conflict will occur if the same structure is plotted without the linking methylene group (—CH_2—) using the same tracking routes.



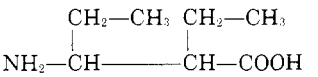
To avoid these situations, the program records and saves the coordinates of the points used in plotting as the structure is being drawn. Whenever a direction change is going to occur, that is, when a branching unit or ring is approached, an inspection of the coordinates table is made. If the plotting is proceeding from left-to-right then the entire area to the right, in both X and Y directions, must be blank before the branch or ring is plotted. If the area is not clear, the bond connecting the branch or ring is extended or "stretched" until the area is available.

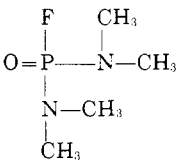
The example above would be drawn as

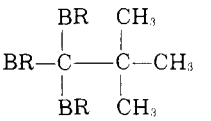


by the program.

The examples shown below have been generated by the display program and have required the bond extension feature to avoid overwriting

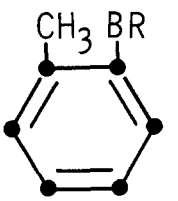




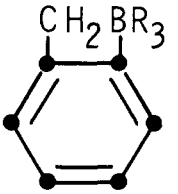


There are some plotting situations which can result in an overwriting conflict but which cannot be remedied by this bond extension technique.

For example, this compound, having a ring with adjacent substituents could be plotted as shown.



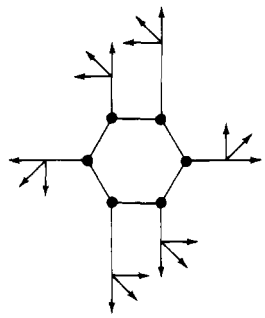
However, if the methyl (CH_3) substituent is replaced by a longer group such as an ethyl group (CH_2—CH_3), then an overwriting error would occur.



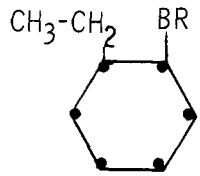
This error cannot be corrected by extending the bonds between ring atoms since the symmetry and shape of a ring must be preserved.

To avoid an overwriting conflict when plotting from a ring it is necessary to define a rigid "out-of-ring" plotting pattern.

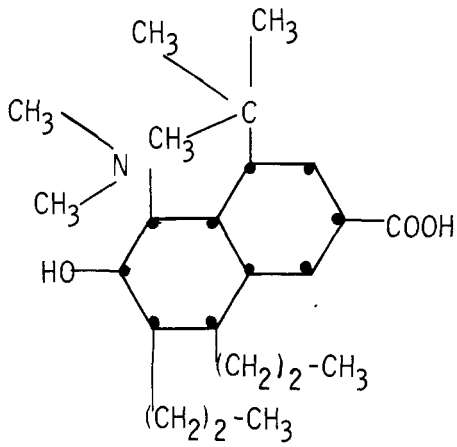
For example, in plotting from a 6-membered horizontal ring the following tracking route pattern is specified.



Using this plotting pattern, the previous example would be drawn by the program as follows:



The example shown below also illustrates the results of plotting according to a specified plotting pattern.



This feature of the display program is important in that it ensures an acceptable structure for a large percentage of compounds, which otherwise may have errors. The overwriting feature is perhaps the most important and at the same time the most complex to write as a computer technique.

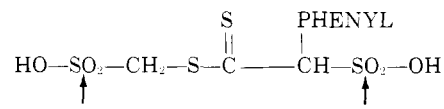
CONVENTIONAL REPRESENTATION OF CHEMICAL STRUCTURES

As well as having the ability to plot a structure without errors, the display program must be able to construct an image which will be familiar to the eye of a chemist.

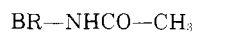
The display program incorporates a large number of detailed statements which are concerned with ways of producing a conventional representation of a structure.

For example, it is conventional to draw certain atomic groups without displaying a bond between atoms. Some examples are given below:

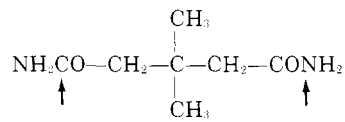
1) SW is ($-SO_2$) NOT ($-S-O_2$) as in,



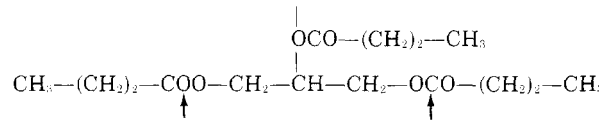
2) MV is (—NHCO—) NOT (—NH—CO—) as in,



3) Similarly, ZV is (NH_2CO-) NOT (NH_2-CO-) as in,



4) and VO is (—COO—) NOT (—CO—O—) as in,



Another convention concerns the representation of atoms which start a molecule. For example, a nitro group is always drawn as $-NO_2$. If the molecule begins with this group, the sequence of atoms is not reversed, even though this implies that the group is linked through the O and not the N.

Consider the following example whose connection table is

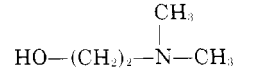
Units W N D D D D N W $NO_2-CH=CH-CH=CH-NO_2$
Connection Transfers
Ring Block

The sequence of units at the start is WN. The translation to atoms would give O_2N- unless special action is taken. The structure shown below has been generated from the connection table and has obeyed this rule of convention.

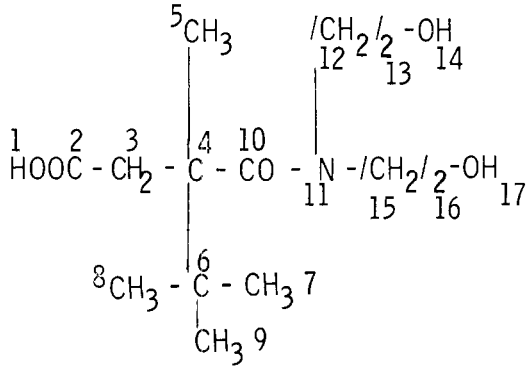
$NO_2-CH=CH-CH=CH-NO_2$

The translation of a unit to its atomic form is usually a simple table look-up step. However, certain units require special treatment if the rules of convention are to be followed. For example, the translation of Q to OH must take into account the cases, when

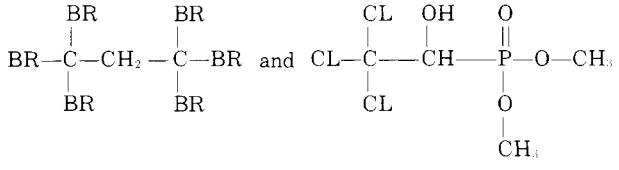
(a) OH appears at the start of a molecule; and if so, it is drawn as (HO—) not (OH—) as in



(b) and where OH appears as a substituent within the molecule; and if so, it is drawn as —OH as in,



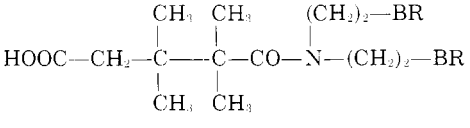
The translation of the unit Q to its atomic form requires special consideration. Other units, such as the halogens E and G do not require this first-unit check. For example,



Finally, the plotting of alkyl chains is another area where rules of convention must be obeyed. These conventions require additional statements in the translation of the unit L to CH_2. The program must take into account the following points:

1) If the unit L is at the beginning or ending of the molecule it must be drawn as CH_3 as in CH_3—CH=CH—CH_3

2) Within a molecule L is represented as CH_2 if linking and as CH_3 when terminal. For example,



where the units are Q V L X L L X L L V N L 2 E L 2 E

linking terminal terminal

3) If the unit L is followed by a number (n) in the units section of the connection, then 2 cases must be considered.

(a) L n is terminal, and the chain must be represented as $(CH_2)_{n-1}$—CH_3 as in CH_3—/CH_2/5—CH=CH—/CH_2/5—CH_3 where the unit string is L6DDL6.

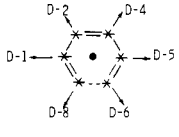
(b) Ln is linking, and the chain must be represented as $(CH_2)_n$ as in NH_3—/CH_2/10—COOH where the units are ZL10VQ.

APPENDIX I

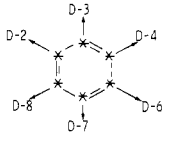
Single Rings

Six-Membered rings

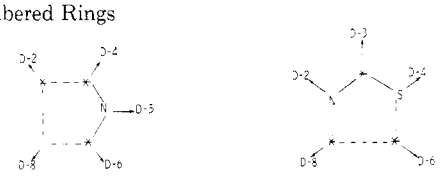
Horizontal



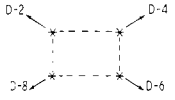
Vertical



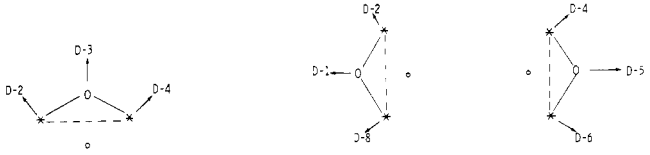
Five-Membered Rings



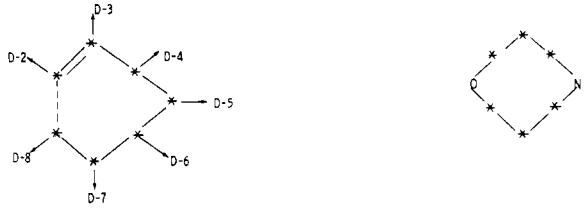
Four-Membered Rings



Three-Membered Rings



Ring Size Greater than Six



Seven-Membered ring

Eight-Membered ring

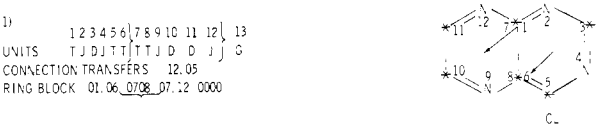
APPENDIX II

Fused Ring Systems

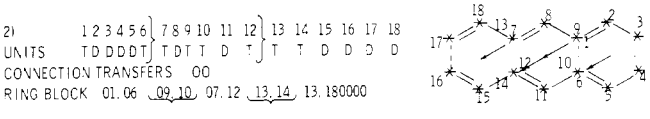
Connection Table

Generated Structure

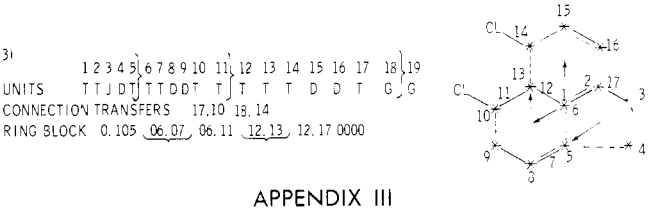
1)



2)



3)



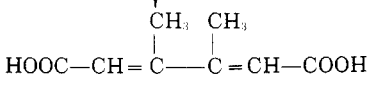
APPENDIX III

Compounds with Double and Triple Bonds

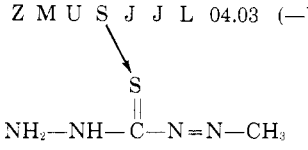
D D L M L D D (D=D—L—M—L—D=D)

1) $CH_2=CH-CH_2-NH-CH_2-CH=CH_2$

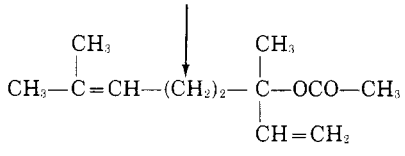
2) 1 2 3 4 5 6 7 8 9 10
Q V D U L U L D V Q 05.04 07.06 (D=U—{—U=D—)



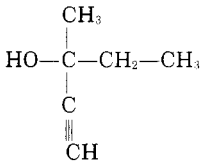
3) 1 2 3 4 5 6 7
Z M U S J J L 04.03 (—U—J=J—)



4) L U L D L 2 X L D D O V L (—U=D—{—X—D=D)
03.02 08.07 10.07



5) Q X L C C L 2
03.02 05.02



LITERATURE CITED

(1) Hyde, E., F. W. Matthews, Lucille H. Thomson, and W. J. Wiswesser, J. CHEM. Doc. 7, 200 (1967).
(2) Thomson, Lucille H., E. Hyde, and F. W. Matthews, J. CHEM. Doc. 7, 204 (1967).

107 Years of Line-Formula Notations (1861–1968)*

WILLIAM J. WISWESSER
Fort Detrick, Frederick, Maryland 21701

Received May 21, 1968

Within seven years after the birth of structural chemistry in 1861, virtually all of the main ideas relating to line-formula conventions were devised and published in the leading chemical journals of a century ago. No basically new practices appeared for some 79 years. Then, within an identically brief period of just seven years (1947–1954), virtually all of the fundamental features of structure-delineating chemical notations appeared in the international chemical literature. The key characteristics of the old conventions and new systems are surveyed.

"The recent international interest in chemical notation has made it seem profitable and desirable to examine the historical records for a guiding background." This was the opening remark for a report on "The History of Chemical Notation" presented 18 years ago at the 118th Meeting of the American Chemical Society in Chicago.[1] That report in turn quoted an opening remark on "Chemical Notation and Nomenclature" by Samuel William Johnson that has an amusing echo today because his remark now is nearly 100 years old:[2]

> "Beginners in Chemistry are liable to much confusion and embarrassment from the fact that there are now in use two distinct systems of Chemical Notation and several forms of nomenclature."

Johnson was referring to notation developments that were then just 10 years old. More recent details on the "Origin of the Line-Formula Method" were given in 1954,[3] and these historic developments again were reported to the American Chemical Society in 1962, this time with a century-old perspective.[4]

The key idea of structural chemistry was popularized in 1860, when the leading chemists of the world attended the first International Chemical Congress at Karlsruhe[5] to resolve their confusions about atoms, molecules, and equivalents. At the close of this 4-day session, Stanislao Cannizzaro clarified the concept of molecules with his reprints on "the message of his old teacher, Avogadro,

* Presented before the Division of Chemical Literature Symposium on Notation Systems, 155th Meeting, ACS, San Francisco, Calif., April 1968.