

## Chemical Notations—A Brief Review\*

By HOWARD T. BONNETT

G. D. Searle & Co., Chicago, Illinois

Received April 11, 1963

The expression "notation" with relation to chemical structures, appears to have become recognized as denoting the representation of a structure by means of a linear series of symbols in a manner which permits a structure to be regenerated. Besides being unambiguous in this way, a notation also gives a unique description of a compound which a code, in general, need not.

A notation has a vocabulary of symbols and rules for grouping the symbols to produce "names." It offers another approach to the problem of "nomenclature."

The development of notations appears to have begun during the 1940 decade. In 1947, the International Union of Pure & Applied Chemistry (hereinafter referred to as IUPAC) appointed a "commission on Coding, Ciphering, and Punched Card Techniques," to examine and report upon ciphering methods.<sup>1</sup> Approximately 10 proposals were offered to the Commission, some of which have never been published.<sup>2</sup> In 1951, the Commission recommended provisional adoption of a notation based on the Dyson cipher.<sup>3</sup> A decade later, considering its work on the formulation of a satisfactory and workable notation system for organic compounds to be finished, the Commission caused the final version of this notation to be published in 1961. In doing so the Commission took the position that the uses and applications by machine methods of its approved notation should be studied later.<sup>4</sup>

Interestingly and significantly, despite these actions of the Commission, others during and subsequent to this interval, have continued to study the notation problem. Thus, Crane and Berry,<sup>5</sup> attempted to combine the Dyson and Wiswesser systems. Silk<sup>6</sup> has continued his studies and has presented a modified notation. Recently, Hayward<sup>7</sup> has proposed a notation designed for computer application, but from which structures can be regenerated.

Other workers<sup>8</sup> have pursued a quite different approach, that of programming the newer machinery to store, search, and reproduce the two-dimensional structural formula. These developments have caused some to wonder if notations have been made obsolete.

The development of notations, and official adoption of one, have led to widespread concern, and even fear among chemists for a number of reasons, one of which has its source in the quite natural and pertinent question, "Why should I learn a notation unless it is sufficiently useful to justify the effort?" One then comes inevitably to the question of utility of a notation. It would seem axiomatic that the success or failure, and the ultimate acceptance of a notation, must stand on its ability to serve in practical applications.

Only recently have reports of attempts to use a notation in an actual operating information system begun to appear. The Army Chemical Center<sup>9</sup> and the Searle Laboratories<sup>11</sup> have reported on the use of the Wiswesser notation. Dyson<sup>12</sup> is installing the IUPAC notation system at *Chemical Abstracts*.

The performance of a notation in a practical application depends upon several factors, including the use, *i.e.*, the nature of the job to be done, the principles underlying the design of the notation, the hardware, and the software available or required to operate with it. In this paper these basic factors will be examined with respect to the IUPAC and Wiswesser systems. In limiting the discussion to these two systems, it is not meant to imply that other proposals, such as those of Silk or Hayward, are not worth examining and testing. Each system requires a book to present its rules, and obviously it is not possible to present a comprehensive review of systems, nor to make an expert of the reader within the span of a single paper. For example, Verkade<sup>2</sup> comments, "The (IUPAC) system is rather complicated at places and can be thoroughly assimilated only after careful study."

As uses of a notation the IUPAC Commission, in its report,<sup>4</sup> mentioned "in indexes, information systems, lexicons of organic chemical data, and in machine systems for correlating the properties of compounds with fractional structural characteristics." Implied in these suggested uses are two completely distinct and exclusive functions to be performed by the notation; namely (1) description, definition or delineation of a structure, and (2) organization of structures for indexing purposes. These are the functions that have been expected of names.

While the structure of a chemical compound represented in notation form may be thought of as a "name," its use in lieu of the usual structural formula, systematic, or trivial name in all the ways those have been used seems questionable. The use of a notation, for example, to convey structures in speech appears less promising than in writing. Even in writing, chemists probably would find notations difficult to follow when substituted for structural formulas in reaction flow charts. Considered in view of the history of nomenclature and of what has been published thus far on the actual use of notations, it is reasonable to conclude that notations will be expected to perform both functions mentioned and that notations probably will find practical application first in the indexing function.

Before leaving the subject of "use," a few general comments are in order. A notation is a new tool. Potential uses merit exploration. These are suitable areas for research.

\* Presented before the Division of Chemical Literature, 143rd ACS National Meeting, Cincinnati, Ohio, January 14, 1963.

Since the functions of delineation and indexing are distinct, it follows that a technique for handling one function may not serve well in handling the other. The structural formula is an example of a technique which does well in delineating a structure, but thus far, remains to be used as a basic indexing tool. Consequently, the designer of a notation system must choose which of the functions he wishes to emphasize. Reevaluation, and perhaps changes, of these choices following operating experience would seem a logical expectation.

The relative emphasis given the delineation and indexing functions in the design of a notation may be influenced by the use and manner of use intended by the designer.

The development of computer equipment has made it possible to examine records in a serial manner, and opens the possibility of subordinating the indexing function. A serial scanning system, however, is not an index.

If a notation is intended to be used for the construction of a manually operable type of index, such as a card file or a list, the principles used in designing the notation must inevitably control the type of organization of structures which will be achieved by alphabetizing the notation. Any system of rules must result in putting certain features of structures into indexing prominence and subordinating other features. Any set of rules will cause the structural features which are subordinated to be scattered throughout such an index, and consequently the index will perform poorly for manual use if these features are the items in which the searcher is interested.

While the principles underlying both the IUPAC and Wiswesser notations cause organic compounds to be divided broadly into acyclic and cyclic structures, they diverge widely in the approach used to benzene and acyclic structures. Benzene compounds are grouped with other cyclic structures in the IUPAC system. Wiswesser treats the benzene ring in a manner analogous to aliphatic structural fragments, because in his view, the benzene ring occurs so frequently as to impair its usefulness as an indexing criterion. According to Wiswesser,<sup>12</sup> the benzene ring occurs as frequently as all other ring systems combined. Wiswesser's view is supported by the Chemical Biological Coordination Center<sup>13</sup> which reported the benzene ring to occur in over 50% of the 53,000 structures in its files at the time.

The rules of the two systems for acyclic compounds give precedence to structural features in the order shown in Fig. 1.

This order of precedence used in the IUPAC system assigns to the most prominent position, *i.e.*, the initial and thus the indexing position, the single letter "C" followed by the number of carbons in the senior component, the side chains, any unsaturation, and the substituent functional groups—all in a purely arbitrary hierarchal sequence. This results in a notation which structurally is analogous to the familiar molecular formula, and for indexing purposes must perform like a molecular formula index.

Wiswesser cites the structural fragments in an end-to-end manner choosing as the starting point that fragment whose symbol is in latest alphanumeric position. Wiswesser notations tend to mirror the structure as drawn; they, too, have some hierarchal characteristics.

IUPAC	Wiswesser
1. Carbon skeleton: initiated by the symbol "C"	1. Longest chain of notation symbols through maximum number of branch symbols
2. Carbon branches; longest cited first	2. End to end citation of symbols in connecting order
3. Unsaturation	3. Starting end is the symbol in latest numero-alphabetic position
4. Functional group substituents in order of precedence <ul style="list-style-type: none"> <li>a. carboxylic acids</li> <li>b. aldehydes</li> <li>c. ketones</li> <li>d. alcohols</li> <li>e. others in symbols alphabetic order</li> </ul>	4. Side chains cited after branch symbols
5. Hetero-atomic, by assembly technique	

Figure 1.

To illustrate the effect of these principles on aliphatic compounds, Verkade<sup>2</sup> selected the compounds shown in Fig. 2. The corresponding IUPAC and Wiswesser notations are included. In the figures which follow, most hydrogen atoms have been omitted from the structural formulas.

In the IUPAC cipher, the carbon chain "C<sub>9</sub>" is first stated, followed by the longest branched chain with its locant "C<sub>4</sub>" and then the methyl branch with its locant "C<sub>6</sub>." Compound II illustrates how the next item, "unsaturation," "E" to represent a double bond, and "Y" to a triple bond, is merely added to the cipher for compound I. Likewise, as shown by the dotted lines, compound III illustrates the further addition to the cipher of functional groups and their locants, *i.e.*, the carboxyl at position 9—"X<sub>9</sub>"; the hydroxyl in position 1—"Q<sub>1</sub>"; and the amino group at position 5—"N<sub>5</sub>." These rules of precedence obviously will group 9-carbon hydrocarbon chains containing an ethyl and methyl side chain and their unsaturated and substitution derivatives, when the ciphers are arranged in an alphabetized list.

In the Wiswesser technique, unbranched saturated carbon chains are represented by numerals, a carbon attached to three atoms other than hydrogen or double-bonded oxygen is represented by "Y," a double bond by the letter "U," a triple bond by "UU," the hydroxyl by "Q," and the carbonyl by "V." These are written in connecting order.

The difference in the performance of the two notations on these three compounds is striking. Whereas the IUPAC notation will group them together, the Wiswesser notation results in three different notations, which do not reveal the identity of the carbon skeleton and would be scattered in an alphabetized list of notations.

In Fig. 3, slight changes of the type commonly made in the pharmaceutical industry, at least, have been made in compound III. Compound IV will be recognized readily as a homolog of compound III and compound V as an analog. Comparison of the IUPAC cipher for compounds III, IV, and V, makes clear that the close relationship of these compounds has been effectively concealed, whereas the Wiswesser notation for these compounds tends to preserve the relationship and to group them in an alphabetized list.

Structural Formula	IUPAC	Wiswesser
$  \begin{array}{cccccccccc}  1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\  C & -C & -C & -C & -C & -C & -C & -C & -C \\  & & &   & &   & & & \\  & & & C & & C & & & \\  & & &   & & & & & \\  & & & C & & & & &   \end{array}  $	C <sub>9</sub> C <sub>2</sub> 4C6	3Y2&1Y3
I		
$  \begin{array}{cccccccccc}  1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\  C & -C & =C & -C & -C & -C & -C & \equiv C & -C \\  & & &   & &   & & & \\  & & & C & & C & & & \\  & & &   & & & & & \\  & & & C & & & & &   \end{array}  $	C <sub>9</sub> C <sub>2</sub> 4C6E2Y7	2UU1Y&1Y2&1U2
II		
$  \begin{array}{cccccccccc}  1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\  HOC & -C & =C & -C & -C & -C & -COOH \\  & & &   &   &   & & & \\  & & & C & & C & & & \\  & & &   & &   & & & \\  & & & C & & NH_2 & & &   \end{array}  $	C <sub>9</sub> C <sub>2</sub> 4C6E2Y7!X9Q1N5	QV1UU1Y&YZY2&1U2Q
III		

Figure 2.

$  \begin{array}{cccccccccc}  1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\  HOC & -C & =C & -C & -C & -C & -C & =C & -COOH \\  & & &   &   &   & & & \\  & & & C & & C & & & \\  & & &   & &   & & & \\  & & & C & & NH_2 & & &   \end{array}  $	C <sub>9</sub> C <sub>2</sub> 4C6E2Y7 X9Q1N5	QV1UU1Y&YZY2&1U2Q
III		
$  \begin{array}{cccccccccc}  3 & 2 & 1 & 5 & 6 & 7 & 8 & 9 & 10 \\  HOC & -C & =C & -C & -C & -C & -C & =C & -COOH \\  & & &   & &   & & & \\  & & & C4 & & C & & & \\  & & &   & & & & & \\  & & & C3 & & & & & \\  & & &   & & & & & \\  & & & C2 & & & & & \\  & & &   & & & & & \\  & & & C1 & & & & &   \end{array}  $	C <sub>9</sub> .(C <sub>5</sub> E1Q3)5C7Y8X10N6	QV1UU1Y&YZY4&1U2Q
IV		
$  \begin{array}{ccccccc}  & & & NH_2 & & & \\  & & &   & & & \\  HO & -C & =C & -C & -C & -C & \equiv C -COOH \\  & & &   & &   & \\  & & & C & & C & \\  & & &   & & & \\  & & & C & & & \\  & & & & & & \text{Benzene Ring}  \end{array}  $	B6:C/6C <sub>5</sub> 4E2Y7X9Q1N5	QV1UU1Y1R&YZY2&1U2Q
V		

Figure 3.

Figure 4 further illustrates this fundamental difference in indexing performance with a group of simple compounds.

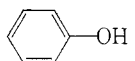
	IUPAC	Wiswesser
C-OH	CQ	Q1
C-C-OH	C <sub>2</sub> Q	Q2
C-C-C-OH	C <sub>3</sub> Q	Q3
C-C-C-C-OH	C <sub>4</sub> Q	Q4
	B <sub>6</sub> Q	QR

Figure 4.

Phenol has been included in this group of otherwise aliphatic compounds to illustrate the treatment of the benzene ring in the two systems. The IUPAC system treats it as a 6-membered aromatic ring coded as "B<sub>6</sub>," in which the ring is given indexing precedence, whereas Wiswesser treats the benzene ring, coded as "R," in a manner analogous to aliphatic structural fragments.

In Fig. 5 are presented the same compounds as in Fig. 4, except that a carboxyl group has been added to the end opposite the hydroxyl group. The figure includes the IUPAC and Wiswesser notations for the compounds

	IUPAC	Wiswesser
$\text{HOOC}-\text{C}-\text{OH}$	$\text{C}_2\text{X1Q2}$	QV1Q
$\text{HOOC}-\text{C}-\text{C}-\text{OH}$	$\text{C}_3\text{X1Q3}$	QV2Q
$\text{HOOC}-\text{C}-\text{C}-\text{C}-\text{OH}$	$\text{C}_4\text{X1Q4}$	QV3Q
$\text{HOOC}-\text{C}-\text{C}-\text{C}-\text{C}-\text{OH}$	$\text{C}_5\text{X1Q5}$	QV4Q
$\text{HOOC}-\text{C}_6\text{H}_4-\text{OH}$	$\text{B}_6\text{X1Q4}$	QVR DQ

Figure 5.

From the notations of these compounds it becomes apparent that the searcher, who is interested in a homologically related series of alcohols, is relatively little worse off in the IUPAC index than he was with the notations of Fig. 4. True, the hydroxyl group has been subordinated to the carboxyl group, but he still is working with an index which resembles a molecular formula index. The searcher in the Wiswesser index is considerably worse off because the end of the molecule having the carboxyl group becomes the end from which the notation starts, with the result that the hydroxyl group becomes subordinated. This illustrates a comment made earlier, that any set of rules must subordinate some functional groups, with the result that an index created by a simple alphabetization of the notation will perform poorly when used manually if the subordinated groups are the subjects of searching interest.

In Fig. 6 is illustrated the use of the IUPAC assembly notation on a group of simple structures in which the carbon chain is interrupted by hetero atoms.

Structure	IUPAC	Wiswesser
$\begin{array}{c} \text{C}-\text{C}-\text{C}-\text{C}-\text{O}-\text{C}-\text{C} \\   \quad   \\ \text{C} \quad \text{C} \end{array}$	$\text{C}_8:2\text{Q}/2\text{C}_2$	3Y&OY
$\begin{array}{c} \text{C}-\text{C}-\text{C}-\text{N}-\text{C}-\text{C} \\   \quad   \\ \text{C}-\text{C} \end{array}$	$\text{C}_6:\text{N}(\text{C}_2)/2\text{C}_2$	3N2&Y
$\begin{array}{c} \text{C}-\text{C}-\text{C}-\text{C}-\text{NH}-\text{NH}-\text{C}-\text{C} \\   \quad   \\ \text{C} \quad \text{C} \end{array}$	$\text{C}_8:\text{C}2:4/\text{N}_2:2/2\text{C}_2$	1Y&MM2Y
$\begin{array}{c} \text{O} \\    \\ \text{C}-\text{C}-\text{C}-\text{C}-\text{O}-\text{C}-\text{C}-\text{C} \\   \\ \text{O} \\    \\ \text{C}-\text{C}-\text{C}-\text{C}-\text{O}-\text{C}-\text{C}-\text{C} \\   \\ \text{O} \\    \\ \text{C}-\text{C}-\text{C}-\text{C}-\text{O}-\text{C}-\text{C}-\text{C} \\   \quad   \\ \text{C} \quad \text{C} \end{array}$	$\text{C}_8:\text{X}/\text{C}_7$	3VO3
$\begin{array}{c} \text{O} \\    \\ \text{C}-\text{C}-\text{C}-\text{C}-\text{O}-\text{C}-\text{C}-\text{C} \\   \\ \text{O} \\    \\ \text{C}-\text{C}-\text{C}-\text{C}-\text{O}-\text{C}-\text{C}-\text{C} \\   \\ \text{O} \\    \\ \text{C}-\text{C}-\text{C}-\text{C}-\text{O}-\text{C}-\text{C}-\text{C} \\   \quad   \\ \text{C} \quad \text{C} \end{array}$	$\text{C}_7/\text{XC}_3$	4OV2
$\begin{array}{c} \text{O} \\    \\ \text{C}-\text{C}-\text{C}-\text{C}-\text{O}-\text{C}-\text{C}-\text{C} \\   \quad   \\ \text{C} \quad \text{C} \end{array}$	$\text{C}_8:\text{C}2:4\text{X}/\text{C}_7\text{C}_2$	1Y&1VO1Y

Figure 6.

For the purposes of assembly notation the IUPAC rules break such structures into carbon chain fragments, called "components," and intervening hetero atoms or groups of atoms called "links." The senior carbon skeleton is cited first, followed by a colon, locant of the senior fragment, link, a stroke, and locant of the succeeding component, etc. The assembly notation is also used on branched carbon structures if the cipher otherwise requires parentheses within parentheses.

The Wiswesser rules permit the X and Y symbols and numerals, representing carbon chain elements, and hetero groups represented by letters to be written in connecting order. The linking hetero groups are readily discernible in both notations, but the rules and resulting notations of the IUPAC cipher are more complicated than the Wiswesser treatment. Both systems abbreviate by omitting symbols under specified conditions as inspections of these notations shows. (In these examples the IUPAC rules omit locants and colons; the Wiswesser omits methyl groups attached to the Y symbol.)

The rules for writing the notation of cyclic structures give precedence to structural features as shown in Fig. 7.

IUPAC	Wiswesser
1. Statement of ring saturation a) A for "saturated" b) B for "aromatic"	1. Type of ring system a) L for carbocyclic b) T for heterocyclic
2. Size and number of rings, largest first	2. Ring sizes and mode of attachment
3. Mode of attachment of rings	3. Ring hetero segments, including keto, in locant order
4. Ring hetero atoms	4. Isolated ring unsaturation
5. Carbon side chains, longest first	5. Statement of ring saturation
6. Isolated ring unsaturation	6. Substituents in locant order
7. Substituent functional groups in order of precedence a) carboxyl b) aldehyde c) ketone d) hydroxy e) others in alphabetic symbol sequence	
8. Implied hydrogen	
9. Hydrogen	

Figure 7.

(With respect to Fig. 7 it should be emphasized that: (1) the items listed are major basic items. Some structures require the use of additional rules, which would be inserted into the sequences shown in Fig. 7; (2) in discussing the Wiswesser notation of cyclic structures, tabulating symbols are used (see p. 33 and 120 of Wiswesser's 1954 manual<sup>12</sup>); (3) the order of precedence given for the Wiswesser system reflects rule revision arrived at on the basis of several years' experience operating indexes based on the notation.)

In denoting cyclic structures (except benzene), both IUPAC and Wiswesser consider the ring system as the senior component. Both use a symbol to signal a ring system. Both use an assembly notation for denoting multiple cyclic structures. Both define hetero segments (with which Wiswesser includes ring carbonyl) after defining the rings and mode of attachment.

The two designers differ widely in their evaluation of the characteristic to use in the first and major indexing position of the notation. In the IUPAC system this position is used to classify a ring system broadly as "aromatic" or "saturated," using "B" or "A," respectively. ("B" is used if the number of double bonds is the maximum number of noncumulative double bonds in monocycles, or equals or exceeds the number of ring carbon atoms not sharing a double bond in polycycles.) Wiswesser uses the initial indexing position of the notation to classify the ring system as carbocyclic or heterocyclic, using "L"

or "T," respectively, for this purpose. Wiswesser uses a pair of symbols "L..J" or "T..J" respectively, to enclose a description of a ring system; the IUPAC system uses an operator signal "Z" to flag ring hetero atoms which are cited in a hierarchal order rather than the locant order used by Wiswesser. Ring substituents are cited by IUPAC in hierarchal order whereas Wiswesser cites them in locant order. Isolated ring double bonds are subordinated to ring carbon chain substituents by IUPAC. Wiswesser includes isolated double bonds within the ring description but subordinates them to hetero atoms.

In Fig. 8 a series of related heterocyclic bicyclic compounds are shown together with their IUPAC and Wiswesser notations.

manual, one may suspect this phenomenon occurs with

some frequency.)

Both the IUPAC and the 1954 Wiswesser manuals cite isomerism and isotopes at the point of occurrence in the notation. The insertion of symbols into the body of the notation, however, inevitably affects the alphabetizing sequence and the appearance of the resulting notations. Thus, in a simple *cis-trans* situation the three versions (*i.e.*, *cis*., *trans*. and unknown or unidentified) may be separated into three, or perhaps four locations in a file. For this reason, in revising the Wiswesser rules, this sort of information is treated as a refinement of the basic structure and is suffixed to the notation.

The IUPAC notation would classify organic compounds into just four groups, namely, "acyclic" designated by "C"; "saturated" ring systems designated by "A,"

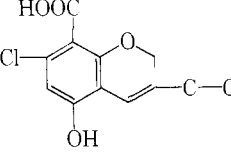
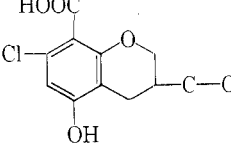
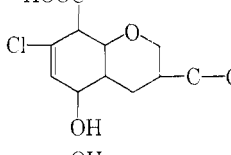
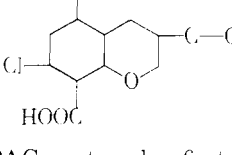
	IUPAC B6.ZQ3C.5CX10Q7Ch9h4	Wiswesser T66 BO CHJ D2 GQ IG JVQ
	B6.ZQ3C.5CX10Q7Ch9h4H56	T66 BOT&J D2 GQ IG JVQ
	A6.ZQ3C.5CX10E8Q7Ch9	T66 BO HUTJ D2 GQ IG JVQ
	A6.ZQ3C.5CX10Q7Ch9	T66 BOTJ D2 GQ IG JVQ

Figure 8.

The IUPAC system by first classifying these heterocyclic compounds on the basis of unsaturation divides the compounds into two major classes, whereas the Wiswesser notation for all is identical through the first six spaces. (This result with the Wiswesser notation is not achieved, however, with the 1954 version of the rules, which also would have placed these compounds in multiple locations in an index. This sort of experience indicated to us that classification of ring systems on the basis of carbocyclic *vs.* heterocyclic character is more useful than on the basis of aromaticity, and also that the ring hetero atoms are relatively more important than unsaturation, and therefore should be given indexing precedence.)

In Fig. 9 the carbocyclic analogs of Fig. 8 are presented. Again, it is noted that the IUPAC system divides the compounds into two major groups on the basis of aromaticity. The two tetralin compounds (III and IV of Fig. 9) further illustrate the interesting effect of design decisions as to relative importance of structural characteristics. The Wiswesser notation gives precedence to the ring nucleus; this results in different locant sets for the substituents. The IUPAC system gives precedence to the substituents; this results in two different "legal" notations for the tetralin nucleus. (From Rule 6.44 of the IUPAC

"aromatic" designated by "B" (which includes benzene), and macrocycles, *i.e.*, rings of ring systems, designated by "M." Wiswesser notations for acyclic and benzene compounds are divided into many groups according to the initiating functional group symbol; other ring systems are classified as carbocyclic or heterocyclic. The number of known organic compounds is estimated to be upwards of 2,000,000. Classification of such a number of compounds into a few main groups in a single index system would result in groups of tremendous size.

Discussion of equipment for use with a notation inherently involves other items also, especially the symbols used in the notation and the programming of the machinery.

The IUPAC notation utilizes upper and lower case letters including several in italics, normal, superscript, and subscript numerals, and in addition, several other symbols. This group of characters, while obtained for use with the equipment at CA on special design and contract from IBM, is available from IBM.

The symbols used for the Wiswesser notation are the 26 letters, the 10 numerals, the ampersand, hyphen, and blank space. These characters are available on accounting equipment.

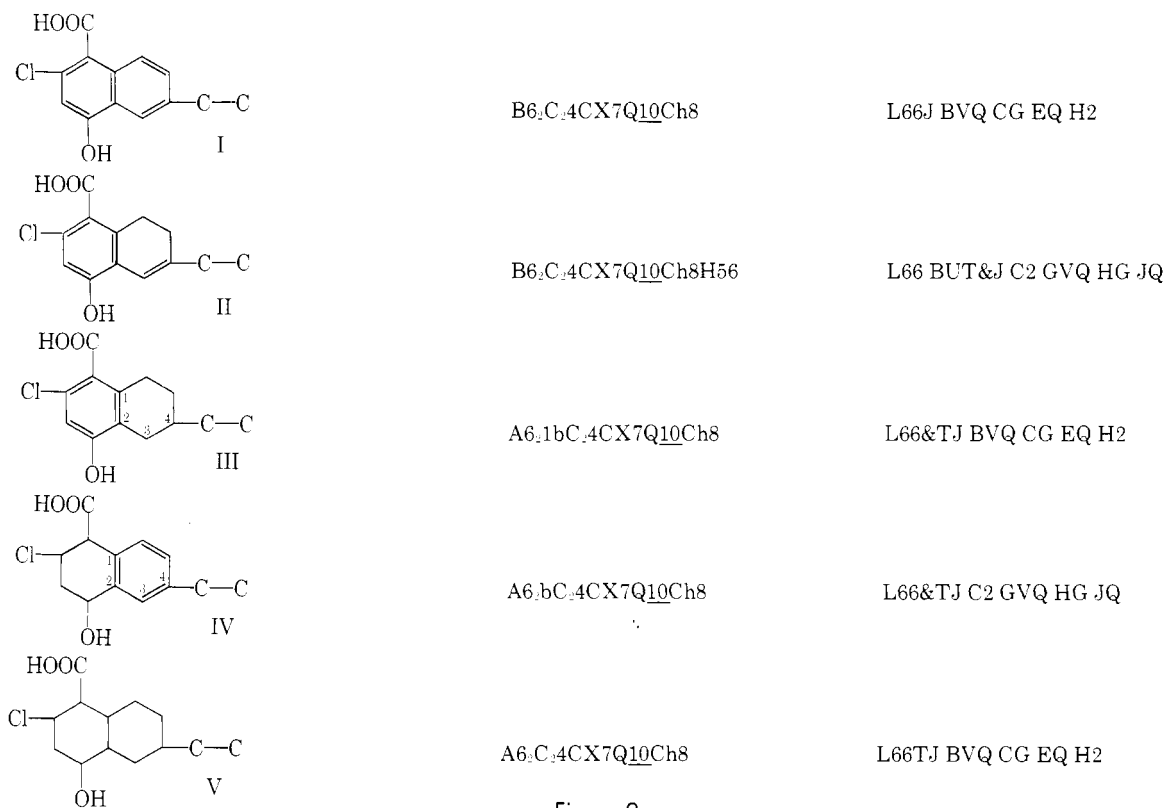


Figure 9.

Most of the characters required by the IUPAC notation and all of those required by Wiswesser could be produced by a standard typewriter, although perhaps not efficiently. Superscripts, for example, could be obtained by manual adjustment of the platen and partial shifting. The 39 characters required for Wiswesser are available on the standard typewriter and require normal typing technique.

Of greater importance is the equipment of the accounting and computing types. Equipment of this nature is required for mechanically manipulating and organizing chemical structures in their notation forms. Generally speaking, readily available standard unmodified accounting equipment offers a limited number of symbols including upper-case letters, the ten numerals, and perhaps as many as a dozen additional punctuation and accounting symbols.

No publications on the use of the IUPAC notation on accounting equipment have appeared other than those relating to the work currently underway at CA. The CA installation includes the following equipment: IBM 026 card punch; IBM 866 typewriter with document writing feature; IBM 1401 computer system equipped with four Model 729 II tape drives and with a special but interchangeable 120-character print-out chain. This equipment contains special modifications which result in an additional rental of several hundred dollars monthly over the corresponding standard unmodified equipment. The special print-out chain is available at a sizeable one-time charge.

In keypunching the IUPAC notation the period, comma, and dollar sign are reserved and used as operator signals to modify the character which follows. A period preceding a letter and numeral indicates a lower case letter and a subscript numeral, respectively. A dollar sign preceding a letter or a numeral indicates a superscript

lower case letter or superscript numeral, respectively. A comma preceding a character indicates the succeeding character is underlined.

The IUPAC manual contains a brief section on arranging ciphers in index form. The process as described appears to be a manual process. It does not include treatment of IBM punch cards containing the operator signals. No procedure for sorting, or instance of sorting IUPAC ciphers with a simple sorter has been published. CA undoubtedly has developed a computer sort program for its application.

In the IBM 1401 computer application at CA, cards containing the IUPAC notations are fed to the computer which converts the punch card notation symbols to two-digit symbols. The double digit form is used for internal core storage and processing, and for storage in magnetic tape. The double digit representation thus increases processing and storage requirements. For print-out, extra logic equipment is required between the processor and printer to convert the double digit symbol to a single character of the print-out chain. This reduces printing speed of the 1403 by over 50%.

The Wiswesser notation has been used at the Army Chemical Center and the Searle Laboratories with the regular keypunches, verifiers, sorters, interpreters, and tabulators of the accounting department. In addition, the Searle Laboratories have used the standard IBM 1401 and IBM 7074 computers.

The symbols of the Wiswesser notation are handled by the accounting equipment in a manner normal for the equipment. Thus, punch cards containing the notations may be alphabetized with the simple sorter, except that the trailer cards required for long structures are held aside and inserted manually. (This chore is eliminated in computer sorting.) The IBM library sort programs for the

IBM 1401 and IBM 7074 computers have been used to alphabetize the notations at high speed. With a 450 card per minute sorter, an operator can alphabetize approximately 500 structures per hour (40 column alphanumeric field). The IBM 1401 with four Model 7330 magnetic tape drives (IBM's slowest model) alphabetized approximately 24,000 random structures in 2.5 hr. The IBM 7074 with six Model 729 IV tape drives alphabetized approximately 37,000 structures in 25 min. plus 5 min. set-up time. Of these the IBM 7074 proved to be the least costly per 1000 compounds.

The limited group of 39 symbols (including the blank space) used in the Wiswesser notation is not an unmixed blessing, since it is necessary to assign multiple meanings to some symbols and to depend upon context for distinction as to meaning. This is analogous to the homonym problem in our language.

Additional symbols other than the slash are represented in IBM punched cards by a three-hole punch pattern. The use of these symbols would complicate the sorting problem and the three-hole symbols would interfere with the alphanumeric array of notations in an index.

Before leaving the subject of machine applications mention should be made of the fact that the power to inspect notations serially opens the possibility of searching for substructures. The group at CA<sup>14</sup> is investigating this possibility and has developed programs for accomplishing such searches. The Searle Laboratories also have developed search programs for use on the Wiswesser notation.

Certain features appear essential to make computer searches practical. Writing separate computer programs for each search could be prohibitively costly. Consequently, the search programs must be in the nature of master programs containing blanks into which the search criteria are written, and it must be possible to run multiple searches simultaneously.

It should be mentioned also that a computer is not necessarily an economic search tool at the point of search. A record or index in magnetic tape or disk form requires a machine to search it. A computer is not necessarily instantly available even within an organization which owns or rents one. The cost of a moderate size computer installation, such as that at CA plus operator, is on the order of \$1.50 per minute. Thus, the relative cost of computer searching *vs.* human searching in an effective index is a suitable area for evaluation. Obviously, it doesn't take many minutes of computer time to be equivalent to an hour of human time. A human can do considerable searching in an hour in an effective index. Here the critical factor is the index—but it is critical for any type of searching. This is not to deny the utility of the computer. It can be an effective instrument in creating indexes for manual use. It can do searches not readily done manually.

The authors of notations, understandably, stress certain factors about their notations. For example, both Dyson and Wiswesser feel that their own notations are more readily recognized than the other. Without denying the existence of relative differences in recognizability, but recalling that the Chinese become accustomed to the Chinese language, the English to the English language, etc., it seems reasonable to suspect that familiarity and

experience are the dominant influences.

Conciseness is an important attribute of a notation. The IUPAC notation has been shortened considerably by the omission of some punctuation marks, which, using the 1958 tentative edition of the manual, accounted for about  $\frac{1}{3}$  of the symbols used. Verkade<sup>2</sup> appears to view the IUPAC notation still to be longer on the average than the Wiswesser. The use of two digit representation of IUPAC symbols in magnetic tape records and within the computer core obviously doubles the storage space required. This factor with the extra print-out time ought to make manipulation of the IUPAC notation relatively costly. The use of blank spaces in the Wiswesser notation makes the notation somewhat longer, but this disadvantage is offset by the advantage that the blank spaces tend to break the notations into "words" which are easier for the human eye to read. If one is operating a card system, he is apt to feel that even the most concise notation is too long.

The authors of notation systems believe their systems to be easily learned. Undoubtedly, the design of a system, the number of rules, etc., will have a marked effect on the ease and accuracy of use of a notation. While the run-of-mill compounds with which most chemists deal can be handled readily by almost any reasonably designed system, I would question that any would be easy to learn and use. For one thing, it is extremely difficult to design rules to handle all foreseeable situations, and to state them with precision and clarity.

Attempts<sup>15</sup> have been made to compare the two notations on ease of learning and accuracy of coding and decoding. Undoubtedly, the design of a notation will have a bearing on these factors. On the basis of our experience in the Searle Laboratories, I am convinced that any individual coding at a reasonable rate of speed and not going back over his work, is going to have a significant level of errors. I am also convinced that even after independently checking the coding effort, a residual error level of 1–2% probably remains unrecognized. We have found this residual error level to obtain in the simple process of copying structures for reproduction in spite of checking by competent people. A comparable error level presumably exists in any code.

In this paper two carefully designed notation systems have been examined with respect to just a few, but basic characteristics. It is clear that the designers of these systems had different objectives in mind. Dyson wants to emphasize the carbon skeleton. Wiswesser wants to emphasize functional groups. The examples given show each approach has its strengths and weaknesses, and these are not coextensive in the two systems. Consequently, by choice of examples, one notation can be made to look relatively better than the other.

It is clear that the hierarchal approach of carbon skeleton plus hierarchal citing of substituent groups used in the IUPAC notation imparts a molecular formula type of appearance and performance to IUPAC notations. The molecular formula index, while it can be arranged in various ways, used in conjunction with other tools, and used with machinery, is nevertheless an index of limited power.

The design of the official IUPAC notation appears to be based on the premise that an effective type of list or

printed index of structures cannot be designed. From this it would follow that the chemist of the future, aside from simple searches, would need special indexes and computer operated indexes. This premise has far reaching implications. Many academic laboratories, many industrial laboratories, and the private individual, all have limited resources, but nonetheless have acute index and search problems in common. It is pertinent to ask whether the premise is either mandatory or valid.

The existence and intensive use of indexes such as those of *Chemical Abstracts* demonstrate utility. The criticisms leveled at them indicate need for improvement. Obviously, changes, whether based on a notation or not, must result in greater effectiveness (and perhaps lower cost) than the molecular formula and systematic nomenclature indexes now available—otherwise what does the new contribute?

From the comments made about the respective notations, the reader may conclude that neither is perfect. Probably both designers would agree. There is no precedent in the history of nomenclature from which one is justified in assuming a notation could be designed that would not be subjected to change.

Naturally, there may be hesitance in using the new. There is a fear of an "orphan" system. Yet, curiously, most information systems in use *are* orphans in the sense that they are unique with their originators and users. They are not compatible, and data in storage is not directly exchangeable among users.

The fear of an orphan system may be exaggerated. Notations, to be successful, must operate with a set of logical rules. A translation program from one to another may be a possibility. Undoubtedly, changes will be made in notation rules which may necessitate extensive revisions in notation indexes. Computer programming, however, can make file changes a practical operation with surprisingly little manual effort. Obviously, change is a serious problem in large printed indexes—although it should be noted that *CA* does make changes, especially at decennial intervals.

Must we wait for a perfect notation? Perhaps the question should be phrased, "Can we wait. . . ?"

The sheer magnitude of the indexing problem will force changes from past practice. A chemist can write notations for relatively complex structures more readily than he can

write correct systematic names. The use of notations could reduce significantly the amount of cross indexing required. Consequently, it is imperative that new techniques be tested and evaluated.

## REFERENCES

- (1) *Comptes rendus de la 14 me Conférence de l' UICPA*, 1947, p. 64.
- (2) "The IUPAC Ciphering System for Organic Compounds," P. E. Verkade *Chemisch Weekblad*, **58**, 137-143 (1962).
- (3) *Comptes rendus de la 16 me Conférence de l' UICPA*, 1951, p. 104.
- (4) "Rules for IUPAC Notation for Organic Compounds," Longmans, Green & Co., Ltd., London, 1961.
- (5) E. M. Crane and M. M. Berry, *Chem. Eng. News*, **33**, 2842 (1955).
- (6) J. A. Silk, *J. Chem. Doc.*, **3**, 189 (1963).
- (7) H. W. Hayward, Research & Development Reports No. 21, 1961, U. S. Patent Office.
- (8) (a) W. H. Waldo, R. S. Gordon, and J. D. Porter, *Am. Doc.*, **9**, 28 (1958); W. W. Waldo and M. DeBacker, *Proc. Intern. Conf. Sci. Info., Area 4*, 49 (1958); A. Opler, *Chem. Eng. News*, p. 108, April 28, 1958; A. Opler and N. Baird, *Am. Doc.*, **10**, 59 (1959); (b) A. Feldman, D. B. Holland, and D. P. Jacobus, Paper presented at 141st National Meeting, ACS, Washington, D. C., March 21-29, 1962; (c) E. M. Crane and Paul Horowitz, Paper presented at 142nd National Meeting, ACS, Atlantic City, N. J., Sept. 9-14, 1962.
- (9) A. Gelberg, W. Nelson, G. S. Yee, and E. A. Metcalf, *J. Chem. Doc.*, **2**, 7 (1962).
- (10) H. T. Bonnett and D. W. Calhoun, *ibid.*, **2**, 2 (1962).
- (11) G. M. Dyson and E. F. Riley, *ibid.*, **2**, 19 (1962).
- (12) W. J. Wiswesser, "A Line-Formula Chemical Notation," Thomas Y. Crowell Co., New York, N. Y., 1954, p. 63.
- (13) E. Dale and K. Heumann, "Statistical Information on Component Parts of Chemical Compounds," Chemical Biological Coordination Center, NAS-NRC, revised March, 1955.
- (14) G. M. Dyson, *ICSU Rev.*, **4**, 110 (1962).
- (15) "Chemical Notation Study, Dyson-Wiswesser. Notation Systems, Encoding Operations, Phase Report" revised, Center for Documentation and Communication Research, Western Reserve University, Cleveland, Ohio, 1960.