Conversion of Connection Table Descriptions of Chemical Compounds into a Form of Wiswesser Notation*

MICHAEL F. LYNCH
Postgraduate School of Librarianship and Information Science, University of Sheffield, England

Received May 21, 1968

The need, in computer files of chemical structures, for a unique linear notation in which the ordering is that of the original notation for manual use is questioned. Instead, a notation using the symbols and syntax of a linear notation and the ordering of a canonical connection table from which it can be derived is suggested for use in mechanized systems. A computer program which converts connection table descriptions of acyclic chemical structures into synonyms of correctly-ordered Wiswesser notations is described. Since they bear a one-to-one relation with the canonical connection table, they are themselves canonical, in a particular sense, and can thus be used for registration, as well as for symbol-connectivity searches. Suggestions are made for extension of the procedure to cyclic structures.

The interconversion of representations of chemical structures has long been of interest to chemists. It dates from the early days of structural chemistry, when it became necessary to verbalize the relations depicted in structural diagrams, and therefore to express in linear form the complexities of branched and cyclic structures. Chemical nomenclature, which has developed since that time in a variety of forms, has been joined by the linear notations, and by representations intended primarily for internal manipulation by computers. The growth of large files of chemical structures in machine-readable form, but in different representations lends urgency to the need for automatic techniques for interconverting these descriptions.

The largest organized files of chemical structures are still in printed, rather than machine-readable form, stored as systematic names in indexes such as those to *Chemical Abstracts* and *Beilstein*, and it is especially heartening that rapid progress is being made toward programs for conversion of nomenclature into connection tables. This is an objective for which algorithms have been discussed by Opler,[15] Tsukerman,[22] and Dyson,[6] and most comprehensively by Vander Stouw.[23] The more limited objective of automatic derivation of a molecular formula was described by Garfield[9] for organic structures, and discussed by Seifer[17] for inorganic compounds. The reverse translation, from connection table into nomenclature, is now also receiving attention, and Conrow[3] has demonstrated the feasibility of this conversion for the class of ring systems for which van Baeyer names are used.

A second important translation is that which converts a structural formula, which is the chemist's most immediate language, into a connection table. This has important connotations in view of the increasing use of computer-typesetting in the production of primary chemical journals, and should result in greater rapidity and lower costs in making structural information available in searchable form. This translation was accomplished by Feldman[8] and by Mullen,[14] using paper-tape chemical typewriters as the input media, and also through optical scanning of hand-drawn chemical structures.[4]

The interconversion of linear notations and connection tables merits close attention today, however, since it is in these representations that the larger files of chemical structures are being built up. It is imperative that facilities for comparison of collections in different structural languages be made available, if duplication of effort is to be avoided. It is clear that of the many translations possible, that from a more highly organized form into a less organized one is the simpler to achieve. We can regard the linear notations as being the more highly organized forms, since they involve complex ordering rules which are applied to obtain the correct canonical notations, while the connection table is the less organized form, although the production of the canonical table is simple and economical. Translation from the IUPAC[16] notation into connection tables was first accomplished in 1963 by Dyson and colleagues,[5] and those for the Hayward[10] and Wiswesser[25] notations have since been programmed.[11,18,19]

The reverse translation, from connection tables into linear notations, is a more complex task, if the prescribed rules of ordering are to be followed faithfully. This necessitates complex flowcharting and programming. Thus far,

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

it has not been completely achieved for any of the major linear notations. However, the rules for generating canonical notations for ring systems in the IUPAC system were successfully programmed and tested on the structures of the *Ring Index* in 1963.[7] More recently, programs for generation of canonical notations for acyclic structures in the Hayward system have been described.[20] The full translations are thus available in one direction only, toward the connection table.

Comparison and interconversion of the contents of files in different representations involves two distinct problems, each with a variety of possible solutions. The first problem is to determine what compounds are common to the two files. When one file is in a linear notation, and the other in the form of connection tables, this can be solved by converting the notations into connection tables and comparing directly, either by producing canonical forms of the connection tables, or by the isomer sort technique. The second problem involves the conversion of connection tables into linear notations.

The Mechanical Chemical Code of Lefkovitz[12] is available as a compact representation which can be used for this purpose. However, many industrial organizations already have sizeable files in the form of notations, and have built up a considerable investment in the form of programs for handling these representations, as in searches for explicit symbols, and production of KWIC indexes. Furthermore, the familiarity of trained staff with the symbology of a notation is a further important asset. The adoption of the MCC could involve a considerable amount of reprogramming.

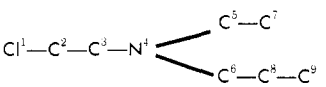
Again, as Lefkovitz[13] has pointed out, manual encodement into linear notations does not reach the ultimate level of confidence in registration that is provided, for example, by the canonical connection table. His solution has been the fact that the MCC takes its uniqueness from the enumeration of the canonical connection table. A further alternative is possible; this is the generation of descriptions using the symbols and syntax of a linear notation, and, as in the case of the MCC, taking its uniqueness from a canonical connection table. The advantages of a representation such as this are considerable, in that it can be handled in computer files by existing programs. Likewise, it can serve as a basis for the generation of fragmentation codes and of KWIC indexes for manual consultation. In only one respect is it deficient, namely, in that a particular notation could not easily be derived manually. However, this limitation affects only direct consultation of a printed index of notations. In the light of the rapidly increasing size of such files and the frequency of their use, and also of work such as that of Thomson *et al.*[21] on the generation of structures from Wiswesser notations on line-printers, it seems likely that manual consultation of notation files will become less important in any case.

A further course of action, which would remove the possibility of noncanonical notations, would be to convert an existing notation file into connection tables, to produce the canonical forms, and then to convert back into the form suggested here. As before, only direct consultation of the file via a manually-produced notation is affected.

With a view to demonstrating the feasibility of an approach to the conversion of connection-table description of compounds into notations, we devised a simple program to accomplish this for acyclic structures. The Wiswesser notation was chosen for the trial, since paths can be described essentially at will through acyclic branches. The approach has certain analogies with the MCC, in that it uses the connection table enumeration as a guide in choosing the path, but differs from it in that the notations which result use the same symbols and syntax as canonical Wiswesser notations; however, no contractions are used other than the use of numerals for unbranched alkyl chains, and no multipliers are included.

The program, written in the list-processing language SLIP,[24] presumes the use of a compact list connection table, in which the enumeration is tree-like, that is, all nodes connected to the first node are numbered successively, then all connected to the second node, etc. It also presumes a connected graph, thus excluding for the moment, molecular complexes.

The program begins by determining the connectivity of each node in the structure and assigns to each a value which is one less than the connectivity. This is determined by totalling the occurrences of each node number in the attachment column, correcting that for the first node by subtracting one. Thus terminal atoms have a value of zero, connecting atoms a value of one, and branching atoms either two or three. This is illustrated in Figure 1 for *N*-2-chloroethyl-*N*-ethylpropylamine.



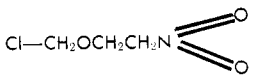
Cl	–	–	–	
C	1	1	1	
C	2	1	1	G2N2&3
N	3	1	2	
C	4	1	1	
C	4	1	1	
C	5	1	0	
C	6	1	1	
C	8	1	0	

Figure 1.

The program then traces a path, beginning at the first terminal atom—i.e., the first atom with a connectivity value of 0—and continuing through the structure until another terminal node is encountered.

As each node is traversed, a symbol for the node is stored. A check is made on the connectivity value; when this equals or exceeds two, this node number is pushed down on a list; when the connectivity value of the node is three, the node number is entered twice. When a node has been traversed, that row in the list is erased. When a terminal node has been encountered at the end of a chain, the topmost node on the push-down list is taken, the first connection for this found, and the process continued.

As each node is encountered, the bond is examined. If a single bond is found no action is taken, but when a double or triple bond is found, the symbols U or UU are entered. However, when an oxygen atom is doubly bonded, the program determines whether it is attached to carbon, when a V is entered, or whether a dioxo group is present, when a W symbol is used. It precedes the attached atom if an oxygen atom is encountered at the outset, but follows it otherwise. Figure 2 illustrates the last case.



Cl	–	–	
C	1	1	
O	2	1	G102NW
C	3	1	
C	4	1	
N	5	1	
O	6	2	
O	6	2	

Figure 2.

As each node symbol is encountered, it is examined to class it by element. There are separate routines for carbon, nitrogen, oxygen, and the halogens. Thus, when a halogen atom is encountered, the appropriate symbol G is entered. If it comes at the end of a chain, no ampersand is inserted.

Carbon atoms are represented by the usual symbols, with the exception that C has not been used; instead, multiple bonds around a carbon atom devoid of hydrogen are stated explicitly by this program. Thus an isolated nitrile group is represented by IUUN. Nitrogen atoms are denoted, as usual, by Z, M, N or K. Figure 3 shows examples of some of these cases.

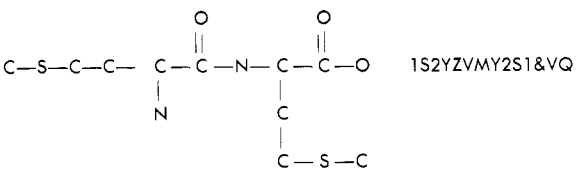


Figure 3.

There is a slight difference, however, in our treatment of the aldehyde group. We do not see the necessity for use of the H symbol, since if the aldehyde group is encountered at the outset, it is rendered by V, if at the end of a branch, by V&, and if the last group in the structure, by V.

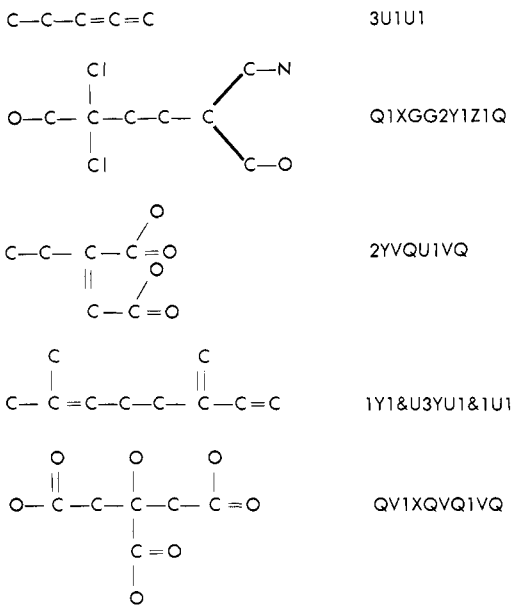


Figure 4.

Some further examples are shown in Figure 4. In each case, the enumeration was started from the left-hand atom. This however, is not a necessary condition for the program, which will deal with any compact list enumeration. In each case, the path chosen through the molecule at branch points is dictated by the lowest number of the next attached atom, thus ensuring a one-for-one correspondence between a particular connection table and the resulting notation.

The program does not at present deal with two-symbol elements other than Br and Cl. We see no difficulty in extending it to these. A further limitation of the program, for the moment, is that it does not deal with the higher valence states of the elements such as P or S in which further ampersands must be inserted to indicate the connections.

RINGS

In contrast to Lefkovitz's MCC, rings cannot be dealt with by direct extension of these symbols. However, we see no problem in extending this work directly to deal with cyclic compounds containing only benzene rings, treating the benzene rings as nodes with variable valence, and inserting letter locants as appropriate. Before considering ring systems other than benzene, however, a fundamental question must be answered. Why do we need canonical notations? The apparent need for canonical notations seems to derive from present-day technology, both from the printed index and from serial files of notations stored in order on magnetic tapes. However, with the increasing use of random-access devices, the need for canonical notations is certainly less acute. It is relevant to mention here recent studies we have made of two factors, firstly, the populations of compounds with specific molecular formulas in a large index, and secondly, the power of simple structural characteristics to partition large molecular formula groups into small subgroups.[2] In the first case, it seems possible, given a particular collection, to predict, on the basis of the molecular formula of a compound presented for registration, the likely population of that molecular formula group. Secondly, given a prediction of the size of the isomeric group, a simple set of characteristics appropriate for that group size can be derived for the particular compound which will result in partitioning of the molecular formula group into very small sets. Lefkovitz has designed the Coded Molecular Formula for this purpose. We have examined the performance of the CMF on some fairly large molecular formula groups and have found that it results in subgroups which contain an average of 1.5 to 3 compounds. We have compared the performance of the CMF with augmented pairs, that is, the atom-bond-atom pairs in which the number of connections at each end of the pair is also given. The results indicate that the augmented pair set lends to improved partitioning, with groups containing on average, 1.1 to 1.2 compounds.

If, by analogy, the symbol strings of the CMF are taken pairwise, then even better resolution can be expected. The same principle can be applied to the symbols of acyclic Wiswesser notations. If this method of registration in on-line systems should become significant, as appears likely, it would seem to suggest that ring descriptions similar to those which Hyde developed from Wiswes-

ser ring notations may be most suitable. Again, since there is undoubted advantage in having ring systems uniquely identified in the notation for search purposes, the Hyde ring descriptions could possibly be developed from canonical connection table descriptions of the ring systems alone.

Further alternative approaches to ring systems exist, of course, including that described in this Symposium by Bowman and his colleagues,[1] or again, table-lookup of canonical Wiswesser ring notations from canonical CT ring descriptions, which might be useful in specialized collections.

To conclude, it would appear that the future provides scope for many alternative schemes for using notations. Hopefully, adequate means of interconverting them can also be provided.

ACKNOWLEDGMENT

Financial support from the office for Scientific and Technical Information, London, and assistance provided by Miss Janet Armitage in the preparation of the program are gratefully acknowledged. Thanks are also due to the National Science Foundation for providing funds for attendance at the Symposium on Notation Systems.

LITERATURE CITED

(1) Bowman, C. M., F. A. Landee, N. W. Lee, and M. H. Reslock, J. CHEM. DOC. 8, 000 (1968).
(2) Bragg, J., M. F. Lynch, J. Orton, and W. G. Town, (in preparation).
(3) Conrow, K., J. CHEM. DOC. 6, 206 (1966).
(4) Cossum, W. E., M. E. Hardenbrook, and R. N. Wolfe, *Proc. Amer. Doc. Inst.* 1, 270 (1964).
(5) Dyson, G. M., W. E. Cossum, M. F. Lynch, and H. L. Morgan, *Information Storage Retrieval* 1, 69 (1963).
(6) Dyson, G. M., *Information Storage Retrieval* 2, 59 (1964).
(7) Dyson, G. M., W. E. Cossum, M. F. Lynch, and H. L. Morgan, "Mechanical Encipherment of Chemical Ring Structures from the Random Matrix," in H. P. Luhn, Ed., *Automation and Scientific Communication* American Documentation Institute, Washington, D. C., 1963.
(8) Feldman, A., D. B. Holland, and D. P. Jacobus, *ibid.* 3, 187 (1963).
(9) Garfield, E., *Nature* **192**, 192 (1961).
(10) Hayward, H. W., *Pat. Off. Res. Dev. Rept.* No. 21, U.S. Patent Office, Washington, D. C., 1961.
(11) Kulpinski, S., *et al.*, 4 "A Study and Implementation of Mechanical Translation from WLN to CT," Vol. 1, Ann. Rept. on Contract NSF C-467, University of Pennsylvania, 1967.
(12) Lefkovitz, D., J. CHEM. DOC. 7, 186 (1967).
(13) Lefkovitz, D., *ibid.* 7, 192 (1967).
(14) Mullen, J. M., *ibid.* 7, 88 (1967).
(15) Opler, A., *Am. Doc.* 10, 59 (1958).
(16) "Rules for IUPAC Notations for Organic Compounds," Longmans, Green and Co., London, 1961.
(17) Seifer, A. L., *Inform. Storage Retrieval.* 1, 29 (1963).
(18) Tauber, S. J., *et al.*, "Chemical Structures as Information," in *Technical Preconditions for Retrieval Center Operations*, B. F. Cheydleur, Ed., Spartan Books Inc., Washington, D. C., 1965.
(19) Tauber, S. J., *et al.*, *Natl. Bur. Stds. Rept.* No. 9587, N.B.S., Washington, D. C., 1967.
(20) Tauber, S. J., *loc. cit.*
(21) Thomson, L. H., E. Hyde, and F. W. Matthews, *ibid.* 7, 204 (1967).
(22) Tsukerman, A. M., and A. P. Terentiev, Proc. Intern. Conf. on Standards for a Common Language for Machine Searching and Translation 1, 493, Interscience Press, 1960.
(23) Vander Stouw, G. G., I. Naznitsky, and J. E. Rush, J. CHEM. DOC. 7, 165 (1967).
(24) Weizenbaum, J., *Comm. ACM.* 6, 524 (1963).
(25) Wiswesser, W. J., "Line Formula Chemical Notation," E. G. Smith, Ed., mimeographed, 1965.

A Chemically Oriented Information Storage and Retrieval System. II. Computer Generation of the Wiswesser Notations of Complex Polycyclic Structures*

CARLOS M. BOWMAN, FRANC A. LANDEE, NANCY W. LEE, and MARY H. RESLOCK
Computation Research Laboratory, The Dow Chemical Company, Midland, Michigan 48640

Received March 26, 1968

A computer program has been written to generate the canonical Wiswesser notation for complex polycyclic structures. The program accepts as input the connection between all the ring atoms and then selects the path which conforms to the notation rules. The operation of the program is described.

A computer-based system has been devised to handle chemically oriented data and information.[2] The system is based on the revised Wiswesser Line Formula Notation.[4] The file organization and methods used to detect notation errors have been described earlier.[1] This paper discusses the difficulties of correctly encoding complex polycyclic structures and a computer program which was written to generate automatically and consistently the canonical notation for these structures.

The notation for a polycyclic structure is obtained by choosing a path through the network which will satisfy

*Presented before the Division of Chemical Literature, Symposium on Chemical Notations, 155th National Meeting of the American Chemical Society, San Francisco, Calif., April 4, 1968.