which is described in a later paper submitted for publication in this journal.

The compacted matrix proved to be an effective record for the generation of open-ended fragment codes. Using simple algorithms to generate an open-ended series of fragments, it is possible to organize files of compounds for both information retrieval and analysis purposes. The resulting files are classified specifically for the problem under examination.

LITERATURE CITED

(1) Dyson, G. M., *Inform. Stor. Retr.* Vol. 1 pp. 66-99.
(2) Gluck, D. J., J. CHEM. DOC. **5**, 43 (1965).
(3) Morgan, H. L., *Ibid.*, p. 107.
(4) Bowman, C. M., *Ibid.*, **7** p. 43.
(5) Wiswesser, W. J., "The 'Dot Plot' Computer Program," Division of Chemical Literature, 152nd Meeting, ACS, New York, September 1966.

Organic Search and Display Using a Connectivity Matrix Derived from Wiswesser Notation*

LUCILLE H. THOMSON, E. HYDE†, and F. W. MATTHEWS
Canadian Industries Limited, Central Research Laboratory, McMasterville, Quebec, Canada

Received July 26, 1967, 1967

A previous investigation of the Wiswesser notation technique for representing chemical structures led to the development of a computer generated connectivity matrix. Having derived a connectivity matrix from the notation, it was necessary to test its suitability for information retrieval purposes. This paper describes the generation of chemical fragments and two-dimensional structural diagrams from the compacted matrix form of the notation.

The investigation had commenced with the objective of carrying out structure/property relationships on organic compound files. Fragment codes are a convenient and economical way of describing a molecule in a file on which mathematical analysis is to take place. If fragment codes were to be used for this purpose, however, it was necessary to have a code specifically designed to reflect the topic under evaluation. Therefore, one application of the connectivity matrix derived from the Wiswesser notation has been to generate fragment codes by algorithms. The first part of this paper describes a program which generates an open-ended fragment code from the connectivity matrix.

The end product of a search of an organic compound file is a list of classified organic structures. Most computer systems in operation today give only a file reference number as the output of a search. A few systems carry a digital representation of the structure, which is available for display either on a computer line printer or a chemical typewriter. Obviously, a computer system which, as output, economically produces structure diagrams is preferable to one giving only file reference numbers. During investigations into various forms of output, consideration has been given to computer generating the structural picture from the search record. The second part of this paper describes a computer program which generates a two-dimensional diagram for chemical structures from the matrix form of the notation.

A COMPUTER GENERATED OPEN-ENDED FRAGMENT CODE

In general, fragment codes break a molecule into recognizable part structures; the fragments chosen depend very much on the nature of the file, and the manner in which it is to be employed.

The object of this work has been to allow a computer to fragment a molecule according to an established set of rules, using as input the matrix form of the notation. The computer generates fragments according to the particular situations met in a molecule; it does not attempt to locate fragments which have been specifically designated. As novel compounds are added to the file, new fragments are generated, and hence the fragment code has the advantage of being open-ended.

The computer program operates directly from the compacted matrix. Each fragment generated is composed of a string of Wiswesser Symbols and varies from two to ten symbols in length, the majority being four symbols long. In general the program reads from a ring or alkyl chain to a terminal group and picks up all symbols.

Wiswesser symbols may be defined and classified according to their connectivity as terminal linking and branching (*1*).

†Present address: Imperial Chemical Industries Limited, Pharmaceuticals Division, P.O. Box 25, Alderley Park, Macclesfield, Cheshire, England

*Presented before the Division of Chemical Literature, 153rd National Meeting, American Chemical Society, Miami Beach, Fla., April 11, 1967.

Table I. Wiswesser Symbols

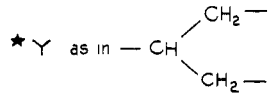
Terminal	Linking	Branching
E, F, G	C, M, O	K, N, P
H, I, Q	S, V	R, S, X
O, M, S		Y
W, Z		

Fragmentation involves the following steps:

1. The program examines the compacted matrix record and locates the branching units. They are classified into two groups:

Group I Branches. Those which can act as starting points for fragments—*e.g.*, rings and alkyls. They are expressed in the notation as:

—R, X, Y*



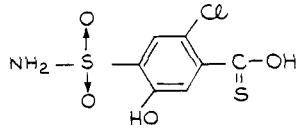
Group II, Branches. Those which are the center of fragments. They are expressed in the notation as:

—K, N, P, S, Y*

*Y as in —CH_2—C(=S)—OH

2. The program then locates all terminal groups (for example, hydroxyl, chloro).

3. In addition, the program is required to generate the longest path in the notation and the points on this path where branching occurs. Consider the following compound:



which is represented by the following compacted matrix record, generated from the Wiswesser notation:

Units ZSWRQGYQS Connection transfers 32, 54, 64, 87
123456789

(a) The branch units in this molecule are at positions 2, 4, and 7. These are tagged so that unit 4 is recorded as a "Group I" unit and unit 2 and 7 as "Group II" units.
(b) The terminal groups are at positions 135689.
(c) The longest path consists of units 12479 and the side branches are 23, 45, 46, 78.

The program reads from the beginning of the molecule and, using this data, develops the following unit combinations, or fragments:

1 2 3 4
4 5
4 6
4 7 8 9

During this operation it was not necessary to examine the Wiswesser units. The routine was performed entirely from the numerical data concerning terminal units, the longest path, and side branches.

As a final step, the four unit combinations listed above are converted into the following fragments expressed in Wiswesser units as:

ZSWR	NH_2SO_2R	
RQ	ROH	R = Phenyl
RG	RCl	
RYQS	RCSOH	

Every fragment is assigned a number, and a compound is registered by entering the compound accession number under each fragment contained in the molecule in an inverted file.

The fragments thus obtained may be listed using a KWIC program. This brings together all fragments containing common Wiswesser symbols. An enquiry made of the file is examined against pertinent sections of the KWIC to establish under which fragments the search should be performed.

One advantage of the KWIC list of fragments is that it will quickly lead the seacher to sub-fragments contained in larger fragments. Hence a situation does not arise where it cannot be decided whether diguanide is a guanide plus, or a guanide is a diguanide minus. This method will always choose diguanide, and by consulting the list of fragments one will extract all fragments which contained the desired symbols. An advantage in preserving the larger fragment is that both the larger and the smaller fragment can be identified readily. If a fragment code requires the smaller fragments to be chosen and expects the questioner to reconstruct the larger, there is always the danger of false coordination. A permuted index of the Wiswesser units in the fragments avoids this difficulty.

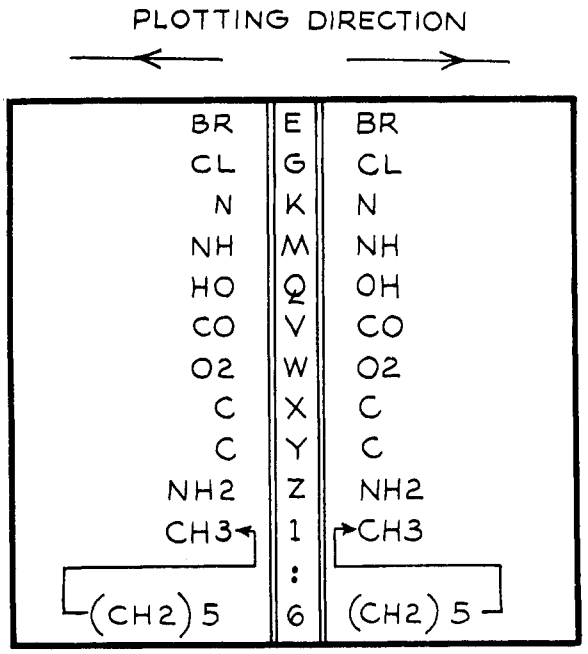
By altering the rules for deriving the "stop" units and varying the definitions of Group I and Group II it is possible to generate different fragments. Therefore, for structure/property relationship, molecules can be fragmented specifically for the problem under examination.

Table II. KWIC Index to Open Ended Fragment Code Derived From Wiswesser Connectivity Matrix

	FRAGMENT	FRAGMENT NUMBER
	*R M V Q	791
	*A M V R	570
	*R M V R	135
	*R M V Z	456
	*A M V Z	059
*A M Y M	M Y M A	476
	*A M Y M M R	693
	*A M Y M M Y M A	476
	*A M Y M M Y Z M	032
	*A M Y S N R R	312
	*R M Y S S Y S N A A	680
*A M Y M	M Y Z M	032
	*A M Y Z M	259
*R S W	M Z	123
*R V	M Z	234
*R	M Z	468

Where A is alkyl and R is any ring atom (Not R as used in Wiswesser Notation)

Table III. Translation of Wiswesser Units



GENERATION OF STRUCTURE DISPLAY FROM A WISWESSER CONNECTIVITY MATRIX

Excellent structure display can be obtained by using a chemical typewriter for computer input, and a line-printer for output (*2, 3*). However, input to a system by chemical typewriter is relatively slow, and results in a long record. The object of this work was to establish the feasibility of using the compact record derived from the Wiswesser notation to generate an acceptable structure display for output on a line-printer. An advantage of this approach is that one record serves the dual purpose of both search and display.

A computer program to generate chemical structure diagrams from the Wiswesser notation has been written and successfully tested. A two-dimensional picture can be generated, using the compacted matrix form of the notation.

However, a program for generating display must compete cost-wise with the alternative method of holding a separate tape record. This limitation must inevitably lead to a compromise. At some point the computer generation of a structure will be more expensive than holding a separate record. In a number of cases a certain amount of difficult draftsmanship is required to generate a structure in an acceptable form. In these cases it is proposed that a separate display record would be created, as a generated display from the connectivity matrix would be impractical.

Generation of a Structure Display from the Wiswesser Connectivity Matrix. The steps involved in converting the compacted matrix to a printed structure are as follows:

Deriving the connection paths for the Wiswesser chemical units.

Translating the Wiswesser chemical units into their normal atomic representation.

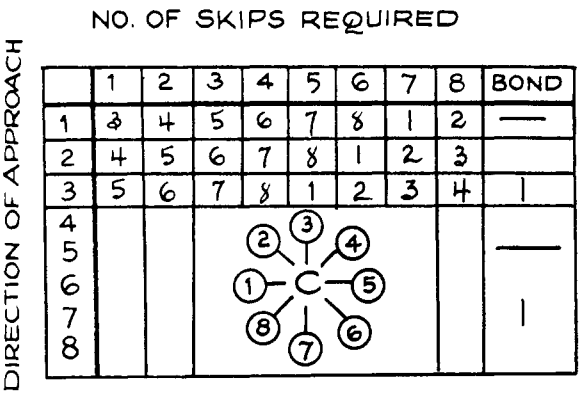
Plotting each atom and bond according to a free plotting routine, which takes into consideration (a) the direction changes required at branching points, (b) the direction changes required to plot points on a ring, and (c) the ability to modify a particular tracking route when an overwriting conflict is likely to occur.

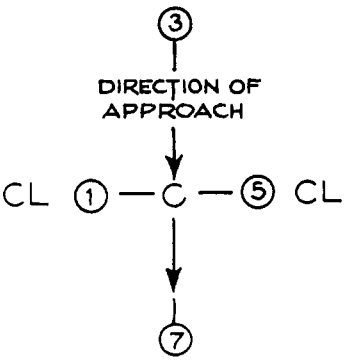
The first two steps deal with the rearrangement and translation of Wiswesser symbols into a format suitable for plotting. The derivation of unit connection paths from the compacted matrix record has already been dealt with in the description of fragment codes. Translation of Wiswesser symbols into normal atom representation requires a look-up in one of two conversion tables—one for forward plotting the other for reverse plotting. The following table gives translation for the most frequently occurring units.

The coordinates of points required to print a structure are derived and plotted in a grid area defined in the memory of the computer. Branching atoms such as N, P, *sec* or *tert* C represent points from which directional changes are required. In plotting from such a point the computer has available eight possible tracking routes, and selects them in a clock-wise order. Having established the direction of approach, and the desired angles required between branches, the tracking routes are obtained from a table which provides all possible directional changes from a particular track.

Considering the example of the *tert* C with 2 Cl side branches:

Table IV

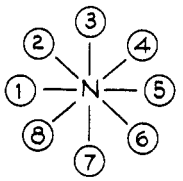




The original path is along direction 3; tracks 5 and 1 have been pre-selected as the preferred routes for plotting the two branches in the directions shown. By referring to direction 3 in the table a change to direction 5 involves skipping over one tracking route, while going to direction 1 from 3 requires a skip over 5 tracking routes. Direction 7 has been pre-selected as the preferred path for the longest branch or continuing portion of the molecule. Going from 3 to 7 involves a skip of 3 tracking routes. Hence the rule for the arrangement shown is skip 1, then 5, then 3. By slight modification to the original matrix record this rule can be applied to all branching atoms irrespective of the number of substituents, and is applicable to all plotting directions by use of the table given. Table V shows all possible directional changes, as well as the print character required as a bond symbol for each tracking route.

Table V. Direction Changes for Acyclic and Cyclic Structures

DIRECTION	NO. OF SKIPS								BOND
	1	2	3	4	5	6	7	8	
1	3	4	5	6	7	8	1	2	—
2	4	5	6	7	8	1	2	3	/
3	5	6	7	8	1	2	3	4	|
4	6	7	8	1	2	3	4	5	\
5	7	8	1	2	3	4	5	6	—
6	8	1	2	3	4	5	6	7	/
7	1	2	3	4	5	6	7	8	|
8	2	3	4	5	6	7	8	1	\



N = BRANCHING ATOM OR CENTRE OF ORIGIN OF RING.

SYMBOLS

• **Backward Slash;** # **Horizontal double bond;** H **Vertical double bond**

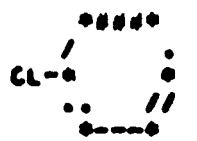
STRUCTURES **Input**

UNITS NDSTTEE

CONNECTION TRANSFER 06040705

RING BLOCK 0105

- - - - - - - -

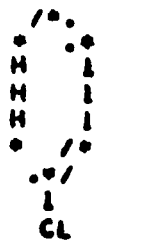


Horizontal

UNITS GTDDDDD

CONNECTION TRANSFER

RING BLOCK 0207



Vertical

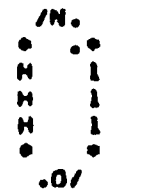
- - - - - - - -

Horizontal

UNITS NDLODD

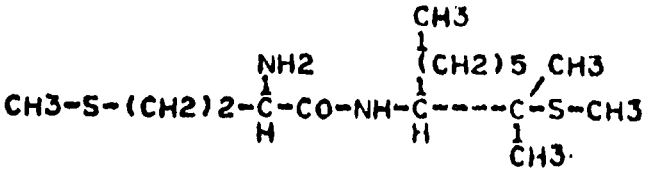
CONNECTION TRANSFER

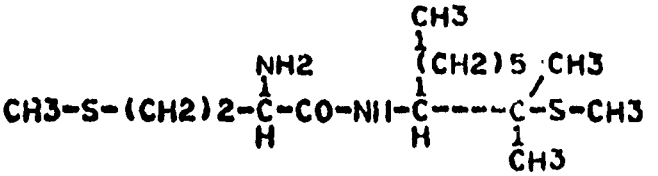
RING BLOCK 0106

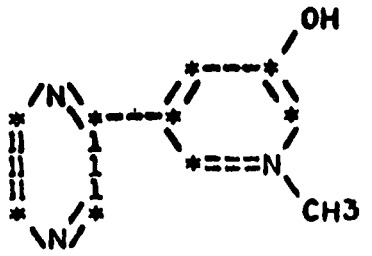


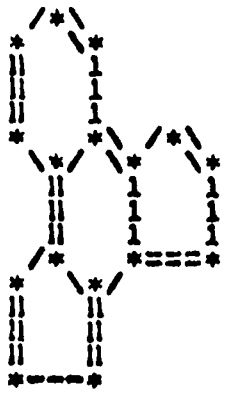
Vertical

Figure 1. Generated structures IBM 1410 line printer









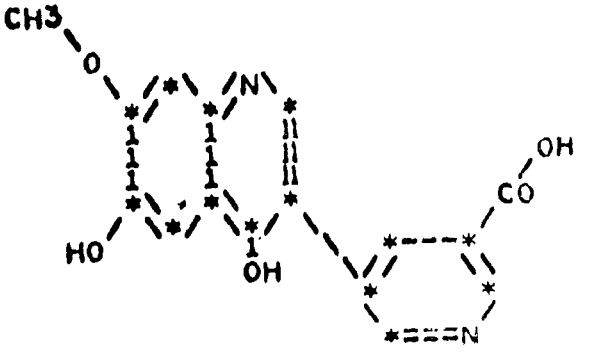
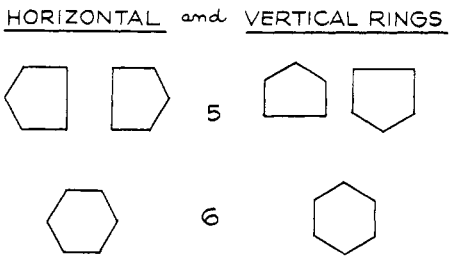
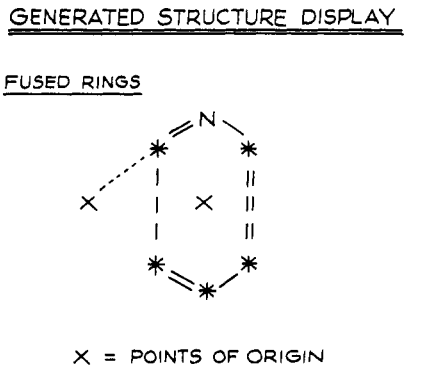


Figure 2. Generated structures printed on ICT 1004

The routine which plots points on a ring considers the ring center to be a branching point with the same eight possible tracking routes. Depending on the size of the ring and direction of approach horizontal or vertical rings are plotted by a selection of the available direction.



In plotting a fused ring system the coordinates of the lowest ring fusion atom in the previous ring are noted and used to calculate the ring center or point of origin of the next ring.



The program generates the ring atoms and the bonds linking them by inspecting the ring portion of the matrix. Heteroatoms M and V are converted to NH and CO, respectively, and carbon ring atoms D, T, L, X, Y are converted to an asterisk mark. All atoms which require to be singly bonded within a ring (*e. g.*, O, S, V, M, L, Y) are tagged. Alternating double bonds are then inserted between the remaining atoms, always commencing with a double bond.

An additional routine notes the lowest and highest coordinates plotted on each line. These data are required only at a branching point or ring center as a direction change might result in overwriting. If a particular path would lead to this overlap problem, the bond leading to the branching point or ring is extended or stretched until the problem area is cleared.

This program has been written for an IBM 1410 with a 48-character line-printer which lacks a number of characters essential for ring display.

It will be converted for use on an IBM 360 linked to an ICT 1004 (Univac), a computer having the required print characters. Figure 1 shows some of the structures obtained in testing the program on the IBM 1410. Figure 2 shows structures also generated by the 1410, but printed on the ICT 1004. The experience gained so far indicates that the concept of structure generation from notation is a feasible one. In many cases, the draftsmanship or acceptable representation of complex molecules would be impractical to generate from any linear record. For these,

it is the intention to create and hold a separate display file using a chemical typewriter for input.

The effectiveness of the matrix record for generating fragment codes and structural display has been established. The consistent choice of symbols and use of logical rules in the Wiswesser notation provide an excellent beginning to a computer program. This notation compacts the chemical and topological data required to represent structures in a way which makes it highly effective in information handling.

LITERATURE CITED

(1) Smith, E. G., *et al.*, "Revised Wiswesser Notation," McGraw-Hill, in press.
(2) Feldman, A., D. B. Holland, D. P. Jacobus, "Automatic Encoding of Chemical Structures," J. CHEM. DOC. **3**, 187, (1963).
(3) Mullen, J. M., "Atom-by-atom Typewriter Input for Computerized Storage and Retrieval of Chemical Structures," *Ibid.*, **7**, 88 (1967).

Documentation of Chemical Reactions by Computer Analysis of Structural Changes*

J. E. ARMITAGE, J. E. CROWE, P. N. EVANS, M. F. LYNCH, and J. A. McGUIRK
Postgraduate School of Librarianship and Information Science, University of Sheffield, Sheffield 10, England

Received May 19, 1967

A method for detection of structural similarities among chemical compounds by computer is described. The method involves an iterative process whereby fragments (of any complexity) common to a pair of structures are generated from their topological descriptions, starting with the atoms common to both and increasing the fragment size one atom at a time until the largest connected set of atoms and bonds common to the pair of structures is determined. Computer programs to perform such an analysis on pairs of acyclic structures are described, and application of the method for identification of structural changes in the reactions of acyclic compounds is discussed.

The provision of easy and adequate access to information on chemical reactions is of fundamental importance to the advancement of chemistry. This area of chemical documentation, however, continues to pose a considerable problem that has not yet been solved even by the application of computers to handling chemical structural information. While the development of systematic nomenclatures and notations, and, more recently, of computer algorithms, has made it possible to identify individual chemical compounds uniquely, no comparable success has been attained with reaction data.

The device most widely used in representing chemical reactions is undoubtedly the reaction scheme or equation, in which the reactants are displayed on one side of the equation, and the products on the other. Such a scheme allows the chemist to deduce the nature of the changes which the molecules undergo, and to see the structural factors which influence the changes. However, the organization of this data, either manually or by computer, is far from simple (*1*). Access through the structures of the compounds which take part in a particular reaction is relatively ineffective, because it gives little indication of the type of reaction involved, nor has any uniform nomenclature of reaction types been developed which is widely accepted and used. Indeed, it is revealing that one of the more effective ways of indexing reactions is by trivial name (*2*)—i.e., by the name of the chemist associated with the discovery of the reaction. The terms "Beckmann rearrangement," "Claisen reaction," "Clemmensen reduction," and "Diels-Alder reaction" are meaningful concepts to every chemist. Descriptive word indexing is less useful, because a simple reaction can be described in many different ways. Thus the rearrangement of cyclohexanone oxime into ε-Caprolactam (Figure 1) can be described variously as an oxime rearrangement, amide formation, lactam formation, or ring enlargement, yet none of these specifies the process uniquely, nor is as adequate as the description "Beckmann rearrangement of cyclohexanone oxime," where the term "Beckmann rearrangement" signifies a pattern of bond rearrangements which is common to the reactions of a large number of compounds.

It seemed imperative that an approach to the documentation of chemical reactions using computer

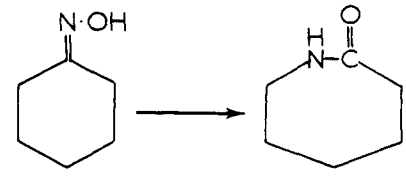


Figure 1.

*Presented before the Division of Chemical Literature, 153rd Meeting, ACS, Miami Beach, Fla., April 1967.