(4) Smith, E. G., *et al.*, "W. J. Wiswesser's Line-Formula Chemical Notation," unpublished data, 1966.
(5) Morgan, H. L., "The Generation of a Unique Machine Description for Chemical Structures—A Technique Developed at Chemical Abstracts Service," J. CHEM. DOC. **5**, 107 (1965).
(6) Lefkovitz, D., and C. T. Van Meter, "An Experimental Real Time Chemical Information System," J. CHEM. DOC. **6**, 173 (1966).
(7) Hyde, E., and L. Thompson, "Organic Search and Display Using A Connectivity Matrix Derived from Wiswesser Notation," Division of Chemical Literature, 153rd Meeting, ACS, Miami Beach, Fla., April 1967.
(8) Eisman, S., A Polish Type Notation for Chemical Structures," J. CHEM. DOC. **4**, 186 (1964).

Conversion of Wiswesser Notation to a Connectivity Matrix for Organic Compounds*

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

Received July 26, 1967

A computer program is described which generates a connectivity matrix using as input an unmodified Wiswesser notation. This program records the topology of a molecule as a statement of the atoms and their connectivity. One symbol is used to represent each atom and this symbol is descriptive of the atom and its bonds. The network of a complex molecule is recorded as a series of interruptions in an assumed linear path. The application of this matrix to information handling of chemical structures is described in a subsequent paper.

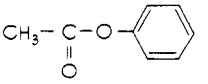
An investigation has been initiated by Imperial Chemical Industries Ltd. to establish a mechanized system for the retrieval and analysis of chemical information. An atom-by-atom connectivity system based on mathematically derived matrixes was considered, but the investigation showed this method to be too cumbersome for the proposed system. Furthermore, in many cases this method destroyed the record of the molecular arrangement of organic compounds.

Having also considered the work reported on the generation of a matrix from the I.U.P.A.C. notation, (*1*), we decided to investigate the usefulness of the Wiswesser notation for this purpose. These investigations have shown that the notation effectively describes the chemistry and the topology required for mechanized retrieval and analysis of chemical information. A computer method has been devised for producing a matrix directly from the notation. This matrix when compacted for tape storage, constitutes a record averaging 60 characters, and is in a form suitable for search and correlation purposes.

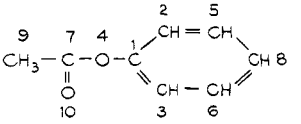
ATOM-BY-ATOM CONNECTIVITY (*2, 3*)

There are two problems associated with any atom-by-atom approach. Firstly, the vast majority of single atoms in any molecule have no descriptive value for search purposes, and secondly an atom-by-atom matrix is a bulky record which is made up of descriptions of atoms and bonds. If the next step is a mathematically generated matrix in order to ensure a canonical ordering of the atoms, then the chemically significant ordering of the atoms is destroyed in the resulting element listing.

The following example will give a clear picture of the disruption of the record of a simple molecule.



The canonical ordering of the atoms derived on a mathematical basis for ultimate magnetic tape storage is as follows:



Thus the record states:

Atom No.	Element	Bond	Connection
1	C	-	-
2	C	L	1
3	C	L	1
4	O	1	1
5	C	L	2
6	C	L	3
7	C	1	4
8	C	L	5
9	C	1	7
10	O	2	7

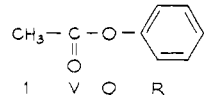
Ring Closure 8–6
L = Alternating bond.

*Presented before the Division of Chemical Literature, 153rd Meeting, ACS, Miami Beach, Fla., April 1967.
†Present address, Imperial Chemical Industries Limited, Pharmaceuticals Division, P.O. Box 25, Alderley Park, Macclesfield, Cheshire, England.
††Present address, U. S. Army, Fort Detrick, Frederick, Md.

NOTATIONS

There is ample evidence that a significant advantage of notations is that they provide an extremely compact method of describing compounds in a computerizable form. Users of the Wiswesser notation have shown rapid coding and economical compound registration (*4*).

Using the Wiswesser notation, the example compound given previously would become



The notation has overcome a number of the problems of an atom-by-atom system. It is canonical in the linear ordering of the notation symbols and this ordering has not destroyed the arrangement of the atoms in the molecule. It is concise because bonds and atoms have been compacted into one symbol, and because of the linear arrangement, there is no need to state connectivity. Finally, it has enriched certain atoms to the point where their chemical significance and differences are clearly shown. The carbons in the example are described as 1 in the methyl group, V in the carbonyl and R in the ring atom. Thus scrutinizing a molecule by computer becomes a much simpler task, as the symbols act as a fragment screen.

However, in achieving these linear representations of molecules, the resulting cyphers are unintelligible except to those who have learned the notation.

CONNECTIVITY DERIVED FROM NOTATION

When retrieving data from an organic chemical file the questions are usually composed of part structures. They require the searcher to retrieve two or more atoms connected in a specific manner. Thus it is of prime importance to show the functional differences of elements and the way they are bonded to other elements as quickly and as effectively as possible. A notation contains the data, but in complex molecules not in an immediately accessible form. It was logical, therefore, to examine the possibility of deriving a connectivity matrix from a notation, and in so doing preserve the advantages of notations outlined above. If this could be carried out, then the concise canonical form provided by the notation would provide a compound registry as a very desirable by-product.

The investigations into devising a connectivity matrix were commenced with multi-substituted benzenes, and it was readily shown that very simple routines were sufficient to obtain a suitable matrix. The work then was extended to notations which cover homo and heterocyclic rings, with side and peri fusions. Thus the investigations have covered a wide range of organic compounds.

DEVELOPING A MATRIX

A computer program devised for converting a standard Wiswesser notation into a connectivity matrix must perform the following two steps.

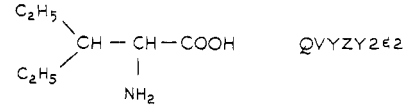
Identify the chemically significant symbols in the notation.

Recognize and label those symbols which either interrupt, or cause a break in an assumed linear path, i.e. branched and terminal symbols.

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						

The following acyclic example will illustrate these steps:



The chemically significant units are:

Q	V	Y	Z	Y	2	2
1	2	3	4	5	6	7

The terminal units (other than first and last units) are: 4, 6.

The branching units are at 3, 5.

Thus the molecule can be represented as a connectivity matrix as follows:

1	Q	$						
2	V	$	$					
3	Y		$	$	$			
4	Z			$				
5	Y				S	$	$	
6	2					$		
7	2						$	

$ = Connection by any bondage.

This array shows that the double diagonal, indicating connections, was interrupted at unit 4 and recommenced at 5 (or expressed simply, the second $ signs opposite 4 and 6 were transferred to 3 and 5, respectively). These interruptions can be described as connection transfers and written in pairs stating the terminal unit first followed by the branching unit. (*e.g.*—43; 65). The example molecule can be represented as:

Units. Q V Y Z Y 2 2
Connection Transfers. 43; 65.

"DOT PLOT" SYMBOLS

The Wiswesser notation does not spell out every single atom in a molecule, but instead points out the type, shows repetition, and indicates change. It is therefore necessary to generate from the notation all excluded atoms, because these constitute nodes in any derived connectivity network. This condition is mainly met in ring systems. For example, the notation for naphthalene is L66J from which is inferred that the compound is composed of two fully unsaturated carbon rings fused together. If it had been other than this the notation would have made suitable notes to this effect. Thus quinoline would be T66 BNJ. The T indicates a heterocyclic ring system and the BN indicates that the carbon in the B position has been replaced by a nitrogen atom.

If a connectivity network is to be composed, then some symbols must be used which do not appear in the notation record.

In earlier work one author of this paper, W. J. Wiswesser, had been working on an entirely different approach for describing ring systems. This system "Dot Plot" comprised spelling out every node in the rings using the following symbols for ring carbon:

L —CH_2—

Y —CH— (As in the notation)

X —C— (As in the notation)

D —CH=

T —C=

The above letters had been chosen carefully so that they would not interfere with existing symbols in the notation. These symbols could be used to expand the ring notation and provide the nodes essential for a connectivity network.

The problem remaining was therefore to examine the possibility of deciphering a standard notation and to generate the above symbols for the omitted portions of the ring record.

GENERATION OF "DOT PLOT" SYMBOLS FROM WISWESSER NOTATION

A program has been written which builds a connectivity matrix using both Wiswesser notation symbols and Wiswesser "Dot Plot" symbols (5). This program is better understood by the consideration of actual examples.

Pyridine  T6NJ

The program detects the number following the T symbol and allocates a linear record of that number of D symbols.

D	D	D	D	D	D
1	2	3	4	5	6

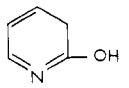
The next step is to read the N, which indicates a nitrogen with no hydrogen attached, and the program overwrites the first D with an N

N	D	D	D	D	D
1	2	3	4	5	6

Thus the matrix for pyridine would be

N	*	$			
D		$	$		
D			$	$	where D is —CH=
D				$	$
D					$
D	*				$

If this compound had contained a substituent—e.g.:

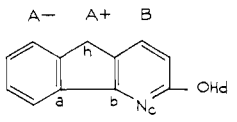


then the notation would be T6NJ BQ. The program reading the BQ adjusts the D at the second location, and the units of the matrix become N T D D D D Q, where T is

—C=

1	N	*	$				
2	T		$	$		$	
3	D			$	$		Connection Transfer
4	D				$	$	6.2.
5	D					$	$
6	D	*				$	
7	Q					$	

If fusions were involved as in the following compound

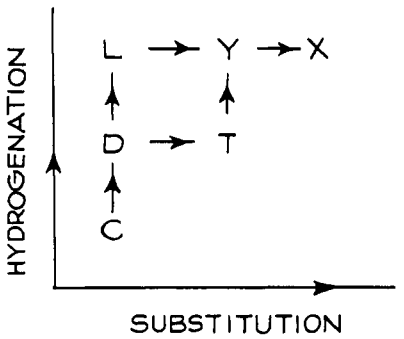


then the notation would be T B656 CN HHJ DQ.

The program works character by character through the notation and commences by examining the first ring which is the one whose lowest character is B, and then the adjacent ring (A+), and finally ring (A−). The record generated would be:

Rings	B	D D D D D D	Modified to	T N T D D T
	A+	D D D D D		T T T L T
	A−	D D D D D D		T T D D D D

CN and DQ would modify the appropriate characters as in the earlier examples and HH, showing that an additional H at the H locant causes D to be replaced by L to indicate —CH_2— as shown in the diagram below. The program notes the overlapping symbols at the fused positions and modifies a D to a T—i.e., changes these from —CH= to —C=,—and, finally, deals with the substituent OH at the D locant.



CONNECTIVITY MATRIX FORM FOR A TAPE RECORD

Experience over a number of years with search systems designed for generic retrieval has shown that it is essential to record the presence of each individual ring in a fused ring system. The computer generation of the dot plot representation of ring systems from the notation repeats the fusion atoms. Hence each ring is separately and completely specified. To prevent the interference of the duplicated nodes in atom-by-atom searching a special descriptive block for ring systems is required in the computer record.

The connectivity matrix for a molecule is described, therefore, by a tape record which is composed of three sections—(a) Units, (b) Connection Transfers, (c) Ring Block.

In this form,

The units are readily accessible for screening purposes.

The terminal and branching units are recorded so that the network of the molecule can be rapidly reconstructed for generic and atom-by-atom searching.

The ring atoms are clearly identifiable as being in the same ring.

Information on ring size is available.

Position of fusion is indicated.

The computer record for the example given in Figure 1 is derived from the notation in the following manner:

The program first compiles the chemical units. In the case of ring atoms, the "Dot Plot" symbols are generated, and the Q is read directly from the notation. This record is a continuous string of symbols of variable length and is terminated by first recording the total number of units, 18 in the example, followed by the digits 99. Having composed a statement of the nodes present, the program next prepares a statement on the connectivity of these nodes. This statement notes where breaks occur in an assumed continuous path. Unit 17 ends the ring system and the next unit Q is not connected to 17 but to the T at 3. This information is recorded in a four-digit record as 1703. The connection transfer statement is always prepared as four digits, one such record for each break. The statement is terminated by the record 0099. The final step in the program is the identification of the ring atoms. This information is written in two four-digit records for each ring present in the ring system. The first four digits record the first and last units in each ring, and the second four digits record ring fusion if present. When fusion occurs as on the example in Figure 1, then the unit numbers of the entry face of the next ring are given. For the example molecule the program derives the following series of numbers. The punctuation marks have been added for clarity and do not constitute part of the record.

0106,0809.0711,1213.1217,0000.

The four zeros indicate both end of the ring statement and the end of record of the molecule.

The example molecule, therefore, is recorded as a continuous variable length record, separated by the record markers 99 and terminated by 0000, and is

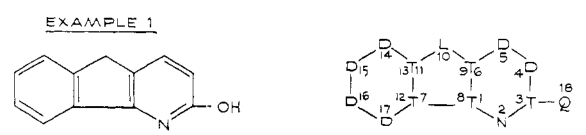
TNTDDTTTTLTTTDDDDQ
189917030099010608090711121312170000

The original notation is not included in the computer record.

The procedure outlined above for face fusion is followed for perifusion, but is modified for spiro rings. In spiro rings the junction unit number is given followed by 00. When two ring systems are present in the molecule then the final four zero record of the first ring system becomes 00 followed by a two-digit number. These two digits record the ring unit number of the unit which is the entry point into the second ring system.

FIGURE 1

EXAMPLE 1



COMPUTER INPUT: WISWESSER NOTATION T B656 CN HHJ DQ

PROGRAM DERIVES: UNITS T N T D D T T T T L T T T D D D D Q
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

CONNECTION TRANSFER: 17 END OF RING, PICK UP CONNECTION AT 3

RING BLOCK: 1 − 6, 8 − 9, 7 − 11, 12 − 13, 12 − 17
SIZE FUSION SIZE FUSION SIZE

DEVELOPING THE MATRIX FROM THE NOTATION

The type of matrix which has been considered essential to convey a rapidly searchable record is derived in exactly the same order as the order of the symbols in the notation. Thus in the fused ring example (1) given earlier there are only 13 ring atoms, but if immediate access to specific rings is required then it is essential to repeat the fused ring atoms as often as they occur in any ring. The 6, 5, 6 of the given notation does precisely this. The order within the notation of ring size, followed by ring atom content, then by state of hydrogenation, and finally by substitution allows the matrix units to be generated and subsequently modified in a manner which leads to efficient program development. From this aspect no symbols have been found redundant, nor has any ambiguity occurred because of insufficient symbols failing in definition.

EFFECTIVENESS OF THE COMPACTED MATRIX

The matrix generated from the notation is a considerable simplification over an atom-by-atom matrix. This is mainly achieved because the notation derives a canonical ordering of the atoms within a molecule and yet leaves them in a chemically sensible arrangement. The linear and numerical form of the compacted matrix leads itself to rapid screening and identification.

As generic searching by fragmentation techniques is by far the most frequent type of searching carried out on chemical files, it was pertinent to examine the usefulness of the compacted matrix for generating fragment codes. To test its efficiency, a fragmentation routine was devised,

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.