(2) J. E. Dubois and H. Viellard, *Bull. Soc. Chim. Fr.*, 905 (1968).
(3) J. E. Dubois and H. Viellard, *Bull. Soc. Chim. Fr.*, 913 (1968).
(4) J. E. Dubois, *J. Chem. Doc.*, **13**, 8 (1973).
(5) J. E. Dubois, *Bull. Chim. Thērapeutique*, 65 (1972).
(6) J. E. Dubois and H. Herzog, *J. Chem. Soc., Chem. Commun.*, 932 (1972).
(7) D. Lefkovitz, *J. Chem. Inf. Comput. Sci.*, **15**, 14–19 (1975).

Structural Search Codes for On-Line Compound Registration

LINDSAY A. EVANS, MICHAEL F. LYNCH,* and PETER WILLETT

Postgraduate School of Librarianship and Information Science, University of Sheffield, Western Bank, Sheffield, S10 2TN, United Kingdom

Received November 8, 1977

A topological index has been developed which can discriminate between isomers in a molecular formula group. This index could be used, in combination with the molecular formula, to provide rapid access to those few compounds in a large chemical structure file which must be compared with a query structure at registration.

1. INTRODUCTION

One of the most common tasks performed by a chemical structure information system is that of registration. This involves determining whether a structure is already present in a machine-readable compound file or if it is new and must be added to the collection. As the compound is entered, it may be given a registration number which serves as a unique link to other information, such as bibliographic references or property data, subsequently added to the file. To ensure that information concerning two, or more, compounds is not to be confused, the structure representation chosen should, ideally, be both unique and unambiguous: a representation is unique if it is the only acceptable one for a given structure while if it describes one and only one possible structure, it is also unambiguous. However, substructure searching, one of the main uses of a compound file, does not require a unique molecular description, so instead of using a canonical representation, any unambiguous representation for a structure, e.g., a connection table derived from an arbitrarily numbered structure diagram, may be used to represent a compound in the file. Registration then involves searching the file for an identical structure; the simpler process of a search for an identical representation is no longer sufficient since a given structure may be represented differently, but equally correctly, when subsequently presented to the system. The atom-by-atom matching[1,2] necessary to establish identity between two differently encoded structures is time-consuming and is too expensive if a large number of structures have to be considered for each registration. To minimize the number of structures which must undergo a detailed investigation, the file is first partitioned into small, nonoverlapping subgroups: only those compounds in the group to which the new compound belongs need to be searched. Registration will therefore become more efficient if the number of compounds in each group is decreased, as fewer atom-by-atom comparisons will be necessary.

On-line registration, carried out at a terminal from which the structure diagram of a compound can be input, requires rapid entry to a direct-access compound file which implies the availability of some form of search code or key, based upon the structure diagram, to obtain access to the collection. Several methods are available for searching disk-based files:[3] one of the simplest is hash-coding[4] in which a key, calculated from the input record, is used to address the disk directly rather than proceeding via some form of directory. Registration would then consist of: input of a structure diagram from a terminal; the automatic generation of a connection table[5] and calculation of the search key; hashing the key to obtain an address in the file; retrieval of all compounds matching the query and display of these compounds on a VDU screen at the terminal. Visual inspection, or atom-by-atom comparison, would then reveal whether the query molecule is already present in the file. To make such a system feasible the number of structures retrieved should be kept as low as possible.

The molecular formula is commonly used to partition the file initially, and this registration method has become known as the "isomer sort" technique; it was used by Ray and Kirsch in the first structure search system.[6] Bragg et al.[7] studied the distribution of compounds appearing in the *Chemical Abstracts* Sixth Collective Formula Index among molecular formula groups and found that the size of molecular formula groups, while highly variable, is also regular, and may be predicted with some accuracy. Methods are therefore needed to partition the larger groups to obtain subfiles small enough to permit atom-by-atom matching. Dyson[8] described a configurational index, based upon the IUPAC linear notation, which indicated the presence of various chemical fragments and could be used in conjunction with the molecular formula index. Shaw[9] found that large molecular formula groups could be partitioned reasonably effectively using a classification of structures based upon the environment of their constituent heteroatoms. Registration with the Mechanical Chemical Code (MCC)[10] was accomplished by a variation of the isomer sort technique; the Coded Molecular Formula (CMF), derived from a symbol count of the atom symbols in the MCC notation, was used instead of the molecular formula, giving smaller groups and hence a more efficient registration process. The extent to which the CMF can distinguish between compounds in the same molecular formula group is discussed by Lynch et al.[11] who characterized each structure in several large molecular formula groups by sets of small, bond-centered fragments generated by an analysis of its connection table. Similar work has been reported by Mishchenko.[12]

While a molecular formula gives a description of the numbers and types of atoms present in a molecule, it does not take into account the manner in which they are interconnected: if this information were available the discriminatory power of molecular formulas would be much increased. The work reported in this paper describes an attempt to provide this information by means of a topological index[13] which condenses the connectivity data present in an adjacency matrix to a single numerical identifier or expression which has (ideally) a different value for every structure.

2. TOPOLOGICAL INDEXES INVESTIGATED

Wilcox[14] has used a simple connectivity index as a measure of molecular branching while calculating molecular π-orbital

0095-2338/78/1618-0146$01.00/0

energies. The index, which will be referred to as WILCOX, is the sum of the squares of the connectivity values of each atom:

$$\mathrm{WILCOX} = \Sigma\, {}^1d_i{}^2$$

where ${}^1d_i$ = connectivity value (number of attachments) of atom *i*.

Randic[15] devised an index χ, to give a numerical value to the degree of branching found in alkane isomers: this takes the form

$$\chi = \Sigma({}^1d_i\,{}^1d_j)^{-1/2}$$

where atoms *i* and *j* are joined by a bond and the summation is taken over all bonds in the molecule. This index will be referred to as RANDIC.

Kier[16,17] has used Randic's index, and higher order indexes based on it, to correlate molecular connectivity with many physical properties and biological activities. The basic Randic index is based upon a single bond; higher indexes are calculated by summing index terms based upon two, three, or more adjacent bonds. Thus:

$${}^2\chi = \Sigma({}^1d_i\,{}^1d_j\,{}^1d_k)^{-1/2},\ {}^3\chi = \Sigma({}^1d_i\,{}^1d_j\,{}^1d_k\,{}^1d_l)^{-1/2}$$

The second-order index,[2] χ, which will be referred to as KIER, therefore represents a summing of three atom fragments where the sum is taken over every possible pair of adjacent bonds. Similarly ${}^3\chi$ is a sum of terms for each and every three-bond path through the molecule, each term being the reciprocal square root of a product of four ${}^1d_i$ values. With such higher order indexes, the amount of computation necessary to identify all the *n*-bond paths becomes very large. An alternative method of deriving higher order indexes was therefore sought and a method based on the extended connectivity values calculated in the Morgan algorithm was used.[18] This is an iterative process of calculating the *n*th order connectivity value of an atom from the (*n* − 1)th order connectivity values of adjacent atoms: it has the great advantage that successively larger structural fragments are described without an actual investigation of the environment around a given atom and is thus much more efficient in computer time than any method that involves some form of path-tracing.

Second order connectivity values, ${}^2d_i$, obtained by one iteration of the Morgan algorithm were used to calculate a second-order index

$$I_2 = \Sigma({}^2d_i\,{}^2d_j)^{-1/2}$$

this being a sum of terms each representing a fragment consisting of up to eight atoms and seven bonds (for two four-valent atoms). Another iteration of the Morgan algorithm allows the calculation of third-order connectivity values ${}^3d_i$ and hence of an analogous index I_3.

The indexes so far considered are based entirely upon the connectivity pattern of the atoms without consideration of the nature of the atoms involved. These, as has been pointed out by Wipke and Dyott[19] and Shelley and Munk,[20] can be used to increase the differentation possible with the Morgan algorithm. An index $I_{1\mathrm{A}}$ was used,

$$I_{1\mathrm{A}} = \Sigma({}^1d_ia_i \times {}^1d_ja_j)^{-1/2}$$

where a_i and a_j are atom values, arbitrarily chosen as C = 3, N = 5 and O = 7 for the three atom types present in the molecular formula groups studied. We can similarly define another index

$$I_{2\mathrm{A}} = \Sigma({}^2d_ia_i \times {}^2d_ja_j)^{-1/2}$$

Table I. The Largest Group Size Obtained When the Two Molecular Formula Groups, $C_{10}H_{10}N_2$ and $C_{10}H_{10}N_2O$, Are Partitioned by Various Topological Indexes

index	$C_{10}H_{10}N_2$	$C_{10}H_{10}N_2O$
WILCOX	44	67
RANDIC	19	23
I_2	9	11
I_3	7	11
$I_{1\mathrm{A}}$	6	5
$I_{2\mathrm{A}}$	2 (3 pairs)	3
$I_{2\mathrm{AB}}$	2 (1 pair)	2 (3 pairs)
$I_{3\mathrm{AB}}$	1	2 (2 pairs)
KIER	9	–
Group size	115	196

A further refinement can be made by including bond type identifiers, as well as atom and connectivity values, in the summation. We define

$$I_{2\mathrm{AB}} = \Sigma[({}^2d_ia_i \times {}^2d_ja_j)b]^{-1/2}$$

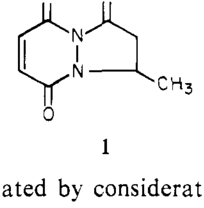
where *b* is the order of the bond connecting atoms *i* and *j*. The final variant tested makes use of all of the refinements discussed so far and is defined as

$$I_{3\mathrm{AB}} = \Sigma[({}^3d_ia_i + {}^3d_ja_j)b]^{-1/2}$$

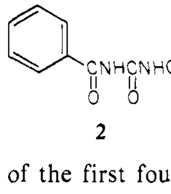
Note that in this case a sum, rather than a product, has been used in the denominator. For the molecular formula groups studied, no differences in performance between $I_{3\mathrm{AB}}$ and the analogous function containing a product were noticed, though this might not be so for other molecular formula groups. The bond type identifiers had the values 1, 2, or 3 for acyclic single, double, and triple bonds; all ring bonds were given a value of 1 (see section 4).

3. EXPERIMENTAL

Seven molecular formula groups were selected as typical of the larger groups, totaling 927 compounds, and varying in size from 70 to 200 compounds per group. The compounds in each molecular formula group were obtained from the Formula Index of the *Chemical Abstracts* 8th Collective Index (1967–1971), omitting polymers, stereoisomers, salts, indefinite compounds, dimers, and adducts. They were then encoded in Wiswesser Line Notation from which CROSSBOW connection tables were generated by a program kindly provided by ICI Pharmaceuticals Ltd.,[21] and these were used to produce redundant adjacency matrices. The various indexes were calculated by user Fortran programs. Single precision arithmetic was found to be sufficient to discriminate between the isomers in each molecular formula group; thus for the group $C_8H_8N_2O_3$ the range of values for the index $I_{3\mathrm{AB}}$ was from 0.655930 to 0.782874 respectively for the two most dissimilar compounds **1** and **2**. Most compounds could be differentiated by consideration of the first four significant figures.



1



2

4. RESULTS AND DISCUSSION

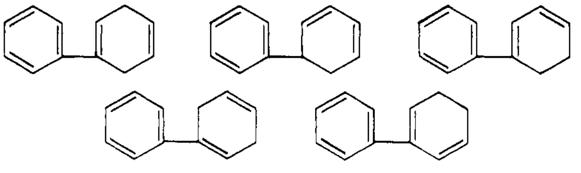
The indexes described above were calculated for each structure in the two molecular formula groups $C_{10}H_{10}N_2$ and $C_{10}H_{10}N_2O$. The degree to which each index discriminated between molecules is indicated in Table I which gives the

Table II. Number of Structures in Each Molecular Formula Group Which Have Identical Index Values

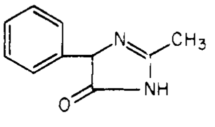
molecular formula	size of group	I_{2AB}	I_{3AB}
$C_{10}H_{10}N_2$	115	1 pair	none
$C_{10}H_{10}N_2O$	196	3 pairs	2 pairs
$C_{10}H_{10}O_2$	200	2 pairs	1 pair
		1 group of 3	
$C_{10}H_{10}O_3$	169	1 pair	1 pair
$C_8H_8N_2O_2$	94	none	none
$C_8H_8N_2O_3$	83	1 pair	none
$C_{12}H_{12}$	70	5 pairs	4 pairs
		2 groups of 3	2 groups of 3
		1 group of 8	1 group of 5

number of structures in the largest unsplit group when the molecular formula group was partitioned by the index values.

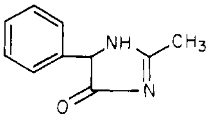
The two indexes I_{2AB} and I_{3AB} were then used to partition five other molecular formula groups: the results are shown in Table II, together with the corresponding results for the groups $C_{10}H_{10}N_2$ and $C_{10}H_{10}N_2O$. The most notable feature is the deterioration in performance when used to partition the smallest molecular formula group studied, $C_{12}H_{12}$. This is almost entirely due to the inability of the connection table generation programs to locate and identify multiple bonds within ring systems. For example, the group of five $C_{12}H_{12}$ compounds not separated by I_{3AB} and shown below would be completely split if ring bond types were taken into account.



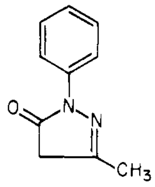
Similarly, the pairs of structures a, b and c, d (from the $C_{10}H_{10}N_2O$ group) would be separable given ring bond information.



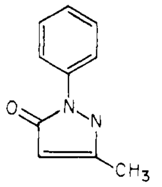
a



b



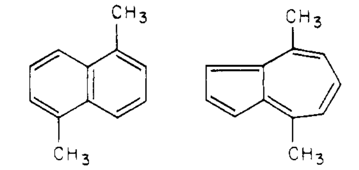
c



d

When the number of heteroatoms relative to the number of carbon atoms is increased, for example, from C_{12} to $C_{10}X_2$ to C_8X_4 or $C_{10}X_3$ to C_8X_5 (where X is O or N), the degree of discrimination achieved between isomers increases. This indicates that the increased differentiation possible between atoms on the basis of atom type outweighs the disadvantage of the smaller number of first-order connectivity values possible (two for oxygen and three for nitrogen), compared with carbon which can have four different values. This is consistent with the fact that there is no noticeable difference in performance when nitrogen atoms are replaced by oxygen atoms, e.g., the groups $C_{10}H_{10}O_3$ and $C_{10}H_{10}N_2O$, and also partially accounts for the relatively low discrimination obtained for the $C_{12}H_{12}$ group.

While carrying out this study we became aware of the work of Freeland et al.[22,23] on augmented connectivity molecular formula (ACMF). Each atom in a structure is assigned an initial value describing the atom type: an augmented connectivity value is then calculated for each atom from the sum of (atom value times bond value) for each of its connections. Higher order connectivity values are then calculated by an iterative process similar to the Morgan algorithm: the process is repeated at least five times and then continued until the number of connectivity values assigned at each step ceases to rise. The ACMF is expressed as a list of atom symbols, each followed by its extended connectivity value, and the list is sorted first by atom symbol and then by connectivity value. The result is not a single number, as for the indexes described in the present work, but the ACMF can, of course, be hashed to give an address. We have calculated ACMF values for all the structures, subject to the ring bond-type constraint: only one pair of compounds was found to be separable by their ACMF's but not by their I_{3AB} values. The structures concerned are:



The difference is presumably caused by the fact that a minimum of five connectivity levels are considered in calculating the ACMF and this indicates that it is very rarely necessary to use higher than third-order connectivity values to discriminate between structures. It is interesting to note that Penny also considered three connectivity levels to be sufficient for his connectivity code:[24] "the use of three levels of connectivity is based upon an intuitive deduction and a superficial survey of the organic compounds."

5. CONCLUSION

A topological index, referred to as I_{3AB}, has been developed which can discriminate to a high degree between isomers in a molecular formula group. This index, with the molecular formula, can be used to identify rapidly those few compounds in a large chemical structure file which must be compared with a molecule input at registration. The majority of isomers receive unambiguous index values, while the largest group of isomers which had identical values consisted of five compounds. The performance could be much improved if different bond values could be assigned to the various types of ring bond: this is a limitation of the structure representation used, not of the technique. The scope of the index could be increased if stereochemical or isotopic data were available. The index has two advantages:

(i) The method of calculation is very simple, since at each step only adjacent atoms need be considered.

(ii) The index can be calculated from a nonunique connection table as its value is independent of the node ordering; it could therefore be used in a WLN-based structure file using a notation-derived connection table.

ACKNOWLEDGMENT

We thank ICI Pharmaceuticals Ltd. for software and Pfizer U.K. Ltd. for the use of their WLN-checker program. L.A.E. and P.W. thank the Department of Education and Science for Information Science Studentships.

REFERENCES AND NOTES

(1) E. H. Sussenguth, "A Graph-Theoretic Algorithm for Matching Chemical Structures", *J. Chem. Doc.*, **5**, 36–43 (1965).
(2) J. Figueras, "Substructure Search by Set Reduction", *J. Chem. Doc.*, **12**, 237–244 (1972).
(3) M. F. Lynch, "Computer-Based Information Services in Science and Technology—Principles and Techniques", Peter Peregrinus Ltd., Stevenage, 1974.

(4) D. M. Murray "A Scatter Storage Scheme for Dictionary Lookup", *J. Libr. Autom.*, **3**, 173-201 (1970).
(5) A. Zamora and D. L. Dayton, "The Chemical Abstracts Service Chemical Registry System. V. Structure Input and Editing", *J. Chem. Inf. Comput. Sci.*, **16**, 219-222 (1976).
(6) L. C. Ray and R. A. Kirsch, "Finding Chemical Records by Digital Computers", *Science*, **126**, 814-819 (1957).
(7) J. H. R. Bragg, M. F. Lynch, and W. G. Town, "The Use of Molecular Formula Distribution Statistics in the Design of Chemical Structure Registry Systems", *J. Chem. Doc.*, **10**, 125-128 (1970).
(8) G. M. Dyson, "Studies in Chemical Documentation", *Chem. Ind. (London)*, 676-684 (1952).
(9) S. R. Shaw, "An Investigation of Some Methods of Improving the Performance of the Molecular Formula in Indexing", unpublished M.Sc. thesis, University of Sheffield, 1973.
(10) D. Lefkowitz, "A Chemical Notation and Code for Computer Manipulation", *J. Chem. Doc.*, **7**, 186-192 (1967).
(11) M. F. Lynch, J. Orton, and W. G. Town, "Organisation of Large Collections of Chemical Structures for Computer Searching", *J. Chem. Soc. C*, 1732-1736 (1969).
(12) G. L. Mishchenko, "Empirical Formulas of Bonds of Compounds and Their Possible Role in Retrieving Factographic Information in Chemistry", *Inf. Probl. Sovrem. Khim.*, 25-38 (1976); *Chem. Abstr.*, **86**, 170058 (1977).
(13) D. H. Rouvray, "The Search for Useful Topological Indices in Chemistry", *Am. Sci.*, **61**, 729-735 (1973).
(14) C. F. Wilcox, "A Topological Definition of Resonance Energy", *Croat. Chem. Acta*, **47**, 87-94 (1975).
(15) M. Randic, "On Characterization of Molecular Branching", *J. Am. Chem. Soc.*, **97**, 6609-6615 (1975).
(16) L. B. Kier, L. H. Hall, W. J. Murray, and M. Randič, "Molecular Connectivity. I. Relationship to Nonspecific Local Anesthesia", *J. Med. Chem.*, **64**, 1971-1974 (1975).
(17) L. B. Kier and L. H. Hall, "Molecular Connectivity in Chemistry and Drug Research", Academic Press, New York, N.Y., 1976.
(18) H. L. Morgan, "The Generation of a Unique Machine Description for Chemical Structures—a Technique Developed at Chemical Abstracts Service", *J. Chem. Doc.*, **5**, 107-113 (1965).
(19) W. T. Wipke and T. M. Dyott "Stereochemically Unique Naming Algorithm", *J. Am. Chem. Soc.*, **96**, 4834-4842 (1974).
(20) C. A. Shelley and M. E. Munk, "Computer Perception of Topological Symmetry", *J. Chem. Inf. Comput. Sci.*, **17**, 110-113 (1977).
(21) E. Hyde, F. W. Matthews, L. H. Thompson, and W. J. Wiswesser, "Conversion of Wiswesser Notation to a Connectivity Matrix for Organic Compounds", *J. Chem. Doc.*, **7**, 200-204 (1967).
(22) R. G. Freeland, S. J. Funk, L. J. O'Korn, and G. A. Wilson, "Augmented Connectivity Molform—a Technique for Recognition of Structure Topology Identity", 169th National Meeting of the American Chemical Society, Philadelphia, Pa., April 6-11, 1975, Abstract CHLT 29.
(23) L. J. O'Korn, personal communication, 1977.
(24) R. H. Penny, "A Connectivity Code for Use in Describing Chemical Structures", *J. Chem. Doc.*, **5**, 113-117 (1965).

The Production of Machine-Readable Descriptions of Chemical Reactions Using Wiswesser Line Notations

MICHAEL F. LYNCH* and PETER WILLETT

Postgraduate School of Librarianship and Information Science, University of Sheffield, Western Bank, Sheffield, S10 2TN, United Kingdom

Received November 8, 1977

A method has been developed for the automatic analysis of chemical reactions by a consideration of the changes in the Wiswesser Line Notations of the reacting molecules. The notations are broken down by a multilevel fragmentation process which yields descriptions for all parts of the molecules. The two fragment lists are compared, duplicates eliminated, and the remaining fragments recombined to produce a reaction site. The output from the program consists of these reaction sites and a set of fragment descriptors derived from them. The method has been tested on a file of 9197 one-reactant, one-product reactions and analyses were produced for 7415 of them (80.6%); the success rate could be increased to ~90%.

INTRODUCTION

Techniques for the automatic retrieval of chemical structural information have now reached both a high level of sophistication and a wide range of applicability;[1] searches may be carried out both for individual molecules and for classes of compounds having certain substructural features in common. The development of comparable means of access to chemical reaction data has proved to be a continuing problem, although the provision of such information is of fundamental importance to the advancement of chemistry. At least part of the problem lies in the multifarious nature of the data, since reaction conditions, yields, mechanism, and the presence of substructural features not actively involved in the reaction may all be of interest. However, the main problem lies in the adequate representation of the reaction site, those parts of the reacting molecules involved in the change, in a machine-readable form.[2]

The most widely used device for representing chemical reactions is the reaction equation, a diagram in which the reactants are displayed upon one side of the equation and the products upon the other. However, the retrieval of reactions by the molecules involved is of limited utility since the main requirement is for substructural transformations such as the conversion of an α,β-unsaturated acid to the corresponding amide or the elimination reactions of dibromo compounds. Vleduts[3] has pointed out that chemical changes generally involve only a limited part of the participating molecules: "A distinctive feature of organic reactions, which involve complicated molecules containing almost exclusively covalent bonds, is the destruction and creation of a comparatively small number of bonds in such a way that, during the process, fairly extensive portions of the molecule do not change their structure." This being so, we should be able to eliminate those parts of the molecules that play no part in the course of the reaction, the remaining partial structures then being taken as describing the reaction sites.

Work in this department has led to two distinct approaches to the automatic identification of reaction sites. In the first[4,5] we sought to map the structures of the reactant and product molecules onto one another so as to identify the largest common fragments and thus, by subtraction, the differences. The work was abandoned owing to program complexity and the amount of processing time required; ten years later, the development of substantially faster computers and of new ways of identifying the common substructures has led us to a reexamination of the potential of such an approach.[6,7] Secondly, we have compared the reactant and product molecules to identify the differences directly; both connection tables[8] and Wiswesser Line Notations[9,10] have been used as the structure representation.

The earlier work[8] used the connection tables for the reactant and product molecules to generate two sets of small, bond-