Advances in Automatic Chemical Substructure Searching Techniques*

W. E. COSSUM, M. L. KRAKIWSKY, and M. F. LYNCH
Research and Development Division, The Chemical Abstracts Service,
The Ohio State University, Columbus 10, Ohio
Received March 23, 1964

The value of a collection of chemical structural information depends in large measure on its versatility in providing for specific and generic searches; that is to say, in how it can be used in locating specific compounds, and in identifying compounds which have certain arrangements of atoms and bonds in common. The latter capability is generally known as substructure search.

Our first substructure searching program, reported at the American Documentation Institute Meeting in Chicago, Ill., in 1963,[1] allowed us to search for substructural features defined as specific groups of atoms, linked by specific bonds. Work with this program showed that while it performed adequately when the queries were spelled out in detail, it lacked generality. For example, the alternatives to specific atoms or bonds in a query were "don't care" atoms or bonds, and not a range of alternatives. Thus, a search for a substructure which included the feature, "a carbon atom joined by a single OR double bond to an oxygen OR sulfur atom," required at least four different specific searches containing:

C—O, C=O, C—S, and C=S.

It was also restricted to contiguous groups of atoms. Thus a request for two or more independent groups in the same molecule required separate searches and correlation of the results.

While we could search for specific rings or combinations of rings, queries which dealt with rings in general terms, for example, "a five-carbon chain in which the second and the fourth atoms are in adjacent rings," could not be answered except by exhaustively stating possible environments. Furthermore, it has been recognized by those working in the field that structure searches involving ring systems pose special problems.

This paper represents the extension of our earlier work, and describes a computer program which permits greater latitude in designating the atoms and bonds of a search, and greater flexibility in posing queries. Both the computer program and the machinable record format are still experimental; full evaluation of their utility will follow in tests on actual search requests. We earnestly ask the reader to submit actual queries for testing.

Representation.—A machine representation of the structural details shown in a two-dimensional graphic formula is needed for mechanical identification of structural features. While we have experimented with the use of the IUPAC notation in substructure search, we feel that a connection table provides the best description of the topology of chemical structures for machine processing. The connection table which we use is a development of that first suggested by Mooers,[2] and tested in search by Ray and Kirsch.[3] In essence, it treats a chemical structure as a graph and describes it by defining the nodes and connectors. The table defines and numbers each nonhydrogen atom and gives the bonds between atoms and the nature of these bonds.

Input to the Connection Table.—The connection table is potentially a language common to a number of input and output methods. We have made preliminary tests of direct clerical generation of the table from a two-dimensional diagram. It can also be generated automatically from linear notations.[5] The Army Chemical Typewriter,[6] the HECSAGON System,[7] automatic optical scanning of hand-drawn structural formulas, and mechanical translation from nomenclature[8] represent other input modes.

Advantages and Disadvantages.—The advantages of the connection table, in brief, are: (1) it can be manually generated, at a reasonable cost, and by relatively untrained personnel, from structural diagrams; (2) it completely represents the topology of a two-dimensional structural formula; (3) it is nonhierarchical, thus no aspects of structure are subordinated to others; (4) it can serve as the basis for the direct automatic computation of the molecular formula for checking purposes; (5) it can be searched by relatively simple programs; (6) it can be transformed automatically into a unique form.

The disadvantages are: (1) it is not compact (some of the advantage gained at search is offset by the length of the record); (2) it is unintelligible to chemists, thus the system must include a parallel store of names, structural diagrams, or notations. However, work is presently in progress to generate graphic structures from the table by machine.

Searches for Acyclic Substructures.—Searches for chains, branched or unbranched, are essentially simple. In the query, the atoms or groups of atoms are represented as numbered parameters. Bonds associated with the atoms, or connecting them, are represented as modifiers to the parameters. Assignment of numbers by the chemist or operator of the system reflects optimizing measures, and, in general, begins with the least common atom or bond of the substructure.

The following are instances which reflect these measures:

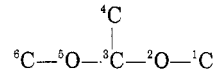
$^{5}C—^{4}C—^{3}C—^{2}C\equiv{}^{1}C$

$^{1}N—^{2}C—^{3}N—^{4}C—^{5}C$

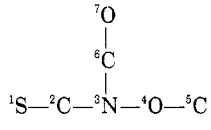
$^{7}C—^{6}C—^{5}C—^{1}As—^{2}C—^{3}O—^{4}C$

* Presented before the Division of Chemical Literature, 147th National Meeting of the American Chemical Society, Philadelphia, Pa., April 1964.

Short branches may be treated by tracing a path twice



in which the list of parameters reads 1234356; but longer branches are best treated as separate groups to reduce retracing through a number of elements

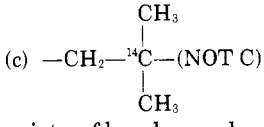


so that the list of parameters is 12345(3)67.

In general, a specific atom, or up to four alternative atoms (with both OR or NOT logic), or a "don't care" atom, can be designated for each atom of the query. Variations in isotopic mass, charge, or valence can also be accommodated. Examples of such searches are:

(a) C—N—C=(O or S)

(b) C=C—C—(Cl, Br, F, or I)



(c) —CH_2—^{14}C—(NOT C)

Further, a variety of bonds may be specified as modifiers to each parameter. These include, at present:

"don't care" bond
single bond only
double bond only
triple bond only
alternating bond (as used to represent an aromatic system) only (referred to in section on ring queries)
single OR alternating bond
double OR alternating bond
single OR alternating OR double bond
double OR alternating OR triple bond

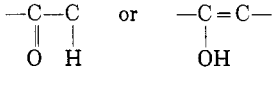
It is also possible to specify an alternating condition so that the possibilities

—N=C—N=C—C

=N—C=N—C—C

=N=C=N=C=C (where = denotes single or double or alternating bonds)

can be coded as a single query. This is particularly useful in searching for tautomeric forms of the same substructure, as with the keto and enol forms of carbonyl compounds



This range of expression contrasts with our earlier program, in which only a specific atom or bond or "don't care" atoms or bonds could be indicated.

Searches Involving Cyclic Substructures.—Providing for more flexible and efficient search for cyclic substructures confronted us with two alternatives. The first was to analyze the connection table mechanically, at input, for certain ring characteristics, and to store these as an integral part of the tape record of each ring structure.[5, 9] Analysis need then be performed only once. The second was to make the analysis at the time of search, and build a temporary record of the features each time the structure was examined in a search. We have chosen the first alternative, on the assumption that rings will be a prevalent feature of searching, and feel that the choice will be validated in actual use. If experience shows that ring analysis at search is more economic, we can revise this step.

Part of the analytical procedure is a search for rings. Each atom in turn is examined to see if it falls within a ring. Once these atoms are identified, the program identifies all rings, but selects as prime rings only those which cannot be divided into smaller rings, and lists them in the tape record. The procedure may result in the inclusion of more rings than are necessary to describe the cyclic system, since alternative routes are considered.

Thus, in the example shown, the cyclic system can be defined by the two rings C—N—C—C—C—O— and C—O—C—C—C—S— alone, but a third possible ring exists, namely, C—N—C—C—C—S—, and this will also be included in the record.

The definition of a prime ring which we have adopted is also used in chemical nomenclature, but whether it adequately reflects chemists' thought patterns must be determined on the basis of its utility. Thus, the norbornane ring is regarded as two five-membered rings, although

many chemists view it as a bridged six-membered ring. Although the machine does not explicitly record rings such as the six-membered ring in norbornane or the ten-membered ring in naphthalene, our search program is fully capable of locating these larger so-called "envelope" rings. The value of the definition will be shown later. However, if its utility is not adequately proved in actual use, it will be modified.

The analytical procedure also examines cyclic bond systems. Input representations, *i.e.*, pictograms and descriptions derived from them, do not distinguish between fixed double and single bonds and those which represent aromatic systems. The program detects those conditions in which interchanged locations of single and double bonds in pictograms of ring systems normally convey the same meaning to chemists, and labels such bonds as alternating.

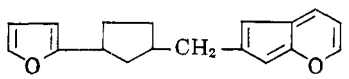
The tape record of each structure, in addition to describing each nonhydrogen atom and its bonds with other atoms, thus lists any prime rings, as defined previously, and includes an identification number for each ring. Appended to the description of each atom in the record is the number of each ring of which it is a member.

Thus, in searching for an acyclic substructure, a proviso can be added to the effect that a specific atom may or may not be a ring member and, further, if a member of a ring, up to three alternative specific ring sizes may be designated (the number is arbitrarily limited by program considerations).

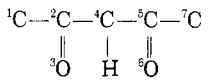
Searches involving prime rings or combinations of such rings of any size are extremely simple. The rings may be designated as being EQUAL TO, UNEQUAL TO,

GREATER THAN, or LESS THAN a specified numeric size.

Examples of various substructural searches that these techniques permit are (a) three five-membered rings not having any common atom, as in

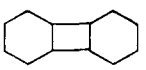


and (b)



where atoms 1, 2, and 4 are members of a ring of more than six atoms, and atoms 5 and 7 are not ring members.

Rings of unspecified size may be searched for, as in: (a) two arsenic atoms in the same ring; (b) N and O in one ring, S in an adjacent ring; (c) O, S, and N in separate rings. In addition, once a ring has been identified during search in response to a request for a ring of unspecified size, that size can be used as a parameter in subsequent logical operations. Thus the logical operators, EQUAL TO, etc., can be applied to rings of previously undetermined size. Examples of the application of this technique are: (a) three rings of the same size; (b) two rings of different sizes joined by a spiro junction; and (c) two rings of any size, separated by a smaller rings, as in



Finally, in searches for both acyclic or cyclic structures, the program has the ability to search for independent, noncontiguous groups, so that compounds containing, for example, the fragment $-NO_2$ and a five-membered ring can be located. The use of both AND and NOT logic between groups further increases the scope of this program.

Search-Limiting Procedures.—The foregoing has been concerned only with searches on individual structure records. This procedure is time consuming and expensive. At least three techniques can be applied to increase the speed of search. These are: (a) screens; (b) search strategy; and (c) file organization.

(a) Screens. These are characteristics of the compound, stored as an integral part of the structure record or generated during search. The query is analyzed for the same features. If the record of these features in the structure file does not correspond to the inclusive or exclusive conditions of the query, the compound is not examined in the atom-by-atom search. The screens which we have used to date are the molecular formula, and the number and sizes of rings, and these are carried in the tape record. In general, however, we feel that exact definition of the screens to be used in a fully operating system must await analysis of a substantial sample of the whole collection, and the nature and volume of the queries addressed to it. The screens, once decided upon, will be generated and revised automatically.

(b) Search Strategy. This involves the statement of the query in such a fashion that its less common characteristics—atoms, bonds, groups, etc.—are placed among the first parameters of the search query, so that decision can be reached as early as possible as to the presence or absence of the characteristic in the stored record. Some instances of this were described.

(c) File Organization. This presumes that it is possible to segment the collection according to criteria which reflect the nature of the queries. Thus, if most searches should include noncarbon atoms, there might be value in segregating the compounds which contain only carbon and hydrogen. A decision on file organization must await analysis of a broad spectrum of actual questions.

Testing.—Small scale tests run on our 1401 computer have shown favorable comparative timings for the second program, but no generalizations of performance can be made yet since the results depend on the choice of queries.

Conclusions.—The flexibility and greater scope of the present search are not gained wholly without losses in other regards. The second-generation program is more extensive than the first and occupies more core storage so that the total number of queries, on average, that can be posed in parallel on our present machine is considerably reduced. The individual potential of the queries, however, is considerably enhanced.

We are aware that many refinements remain to be added, both in the form of the stored record and in the search programs. We intend, however, that in storing the collection of chemical structures in searchable form the format and degree of detail will not be prejudiced by search techniques which are not yet optimal. We will retain the ability to transform the structural record automatically into its most advantageous form. We will seek to accommodate in our future programs the full range of questions that chemists ask today. We will make adjustments once the interactions between chemical thought and selection potential begin to flourish.

Acknowledgments.—We wish to record our thanks to Dr. G. Malcolm Dyson for stimulating discussion leading to this work, and to the Office of Science Information Service of the National Science Foundation for partial support of the study.

REFERENCES

(1) W. E. Cossum, G. M. Dyson, M. F. Lynch, and R. N. Wolfe, in "Automation and Scientific Communication," H. P. Luhn, Ed., American Documentation Institute, Washington, D. C., 1963, pp. 15–18.
(2) C. N. Mooers, *Zator Tech. Bull.*, 59 (1951).
(3) L. C. Ray, and R. A. Kirsch, *Science*, **126**, 814 (1957).
(4) Anon., *Chem. Eng. News*, **41** [49], 35 (1963).
(5) G. M. Dyson, W. E. Cossum, M. F. Lynch, and H. L. Morgan, *Inform. Storage Retrieval*, **1**, 69 (1963).
(6) A. Feldman, D. B. Holland, and D. P. Jacobus, *J. Chem. Doc.*, **3**, 187 (1963).
(7) P. Horowitz, and E. M. Crane, paper presented at the 142nd National Meeting of the American Chemical Society, Atlantic City, N. J., Sept. 1962.
(8) G. G. Stetsyura, and A. M. Tsukerman, *Nauchn.-Tekhn. Inform.*, **3**, 17 (1962).
(9) G. M. Dyson, W. E. Cossum, M. F. Lynch, and H. L. Morgan in ref. 1, pp. 79–82.