Table IV. Comparison Ratios as Percentages Calculated from Target Set 1 and Library Statistics for Three MODN Compressions

	MOD14	MOD10	MOD7
from first target set	8.31	11.00	14.99
from library statistics	7.89	11.23	14.82

fraction of library entries in a partition is used as an estimate of the fraction of target spectra that will require a search of that partition, it is possible to estimate the savings directly from the partitioning statistics of the library file. The sum of the squares of the fraction of entries in each partition gives the ratio of the number comparisons required with file partitioning to the number comparisons required by the conventional search strategy. These ratios, as calculated from library statistics and for the set of 40 target compounds, are reported in Table IV. Because a search algorithm not only compares spectra but also performs tasks such as sorting new matches into a running list of nearest matches and looking up the names of nearest matches, the actual time required for a search of a partitioned file will not be simply the product of the time required for a conventional search and the fraction of the spectral comparisons performed.

It is also interesting to consider the relationship between MODN dimensionality and the use of file partitioning. To do this, the number of comparisons between spectra must be weighted to include the number of dimensions in a spectrum. The comparison of two MOD14 spectra requires the comparison of twice as many pairs of numbers as does the comparison of two MOD7 spectra. Thus, using the conventional search strategy a MOD7 search compares only half the number pairs that a MOD14 search does. However, with file partitioning, the number of spectra in a partition increases as the number of dimensions, and therefore partitions, decreases. MOD14 spectra are still twice as long as MOD7 spectra, but they are distributed among twice as many partitions. For the cases examined the partitioned MOD10 and MOD7 searches require consideration of about 95% as many number pairs as the partitioned MOD14 search. The use of file partitioning makes the search times required for typical MOD14 searches comparable to those required for the lower dimensionality MODN searches.

CONCLUSIONS

The use of MODN spectra offers a promising approach to the compression of mass spectral data for library searching. The MOD14 search probably represents the best choice of the three considering the trade-off between search performance and library storage requirements. The efficiency of the data compression, the molecular structure information implicitly contained in MOD14 spectra, the general performance of the search with target compounds, and the effectiveness of file partitioning are features which recommend use of this data compression method.

ACKNOWLEDGMENT

The authors wish to thank Alan Hanna for helpful discussions during the course of this work. Portions of the work were presented at the 29th Annual Pittsburgh Conference on Analytical Chemistry and Applied Spectroscopy.

LITERATURE CITED

(1) "Index of Mass Spectral Data", AMD11, American Society for Testing and Materials, Philadelphia, Pa., 1969.
(2) S. Abrahamsson, *Sci. Tools*, **14**, 29 (1967).
(3) B. Petersson and R. Ryhage, *Ark. Kemi*, **26**, 293 (1966).
(4) L. R. Crawford and J. D. Morrison, *Anal. Chem.*, **40**, 1464 (1968).
(5) H. S. Hertz, R. A. Hites, and K. Biemann, *Anal. Chem.*, **43**, 681 (1971).
(6) H. W. Brown and E. J. Bonelli, *Abstr. 1977 Pittsburgh Conf. Anal. Chem. Appl. Spectrom.*, 144 (1977).
(7) S. L. Grotch, *Anal. Chem.*, **42**, 1214 (1970).
(8) S. L. Grotch, *Anal. Chem.*, **45**, 2 (1973).
(9) B. A. Knock et al., *Anal. Chem.*, **42**, 1516 (1970).
(10) L. E. Wangen, W. S. Woodward, and T. L. Isenhour, *Anal. Chem.*, **43**, 1605 (1971).
(11) M. C. Hamming and R. D. Grigsby, *Proc. 15th Annu. Conf. Mass Spectrom. Allied Top.*, 107 (1967).
(12) L. R. Crawford and J. D. Morrison, *Anal. Chem.*, **40**, 1469 (1968).
(13) D. H. Smith, *Anal. Chem.*, **44**, 536 (1972).
(14) S. P. Markey, W. G. Urban, and S. P. Levine, Eds., "Mass Spectra of Compounds of Biological Interest", USAEC Technical Information Center, Oak Ridge, Tenn., TID-26553-P1.
(15) R. M. Silverstein and G. C. Bassler, "Spectrometric Identification of Organic Compounds", Wiley, New York, 1967.
(16) S. L. Grotch, *Anal. Chem.*, **43**, 1362 (1971).

Computer-Assisted Examination of Chemical Compounds for Structural Similarities[1,2]

TOMAS H. VARKONY, YOSSI SHILOACH, and DENNIS H. SMITH*

Departments of Chemistry, Computer Science, and Genetics, Stanford University, Stanford, California 94305

Received August 17, 1978

An algorithm for finding common substructures among a potentially large and diverse set of chemical structures is described. The algorithm has been implemented in an interactive computer program called MAXSUB. The program allows the chemist to specify his definition of what constitutes "commonality" of substructures by providing control over the importance of degree of substitution, hybridization, atom type, and ring membership of atoms in substructures and multiplicity of bonds between atoms. Applications to problems involving topological representations of chemical structures, including macrolide antibiotics and marine sterols, are discussed briefly to illustrate the program. Some possible extensions to dealing with three-dimensional representations of structures are mentioned.

Comparison of structural features among a set of chemical structures which display some common behavior is a frequent problem in chemical research. This problem can usually be characterized as one of relating the structures of the molecules to some "activity", i.e., structure/activity relationships in the broadest sense of the term. For example, the activity may be of a physical nature in that the molecules all display some characteristic pattern or subpattern in a spectroscopic technique, or biological in that the molecules demonstrate similar physiological effects. The importance of relating common portions, or substructures, of the molecules to commonly observed activities, whether to build correlation tables of substructures to spectroscopic behavior or to design new drugs,[3] to mention only two applications, hardly needs emphasis. In all such studies it is presumed that molecules displaying similar activities do so because they share some similar feature or

0095-2338/79/1619-0104$01.00/0

features; the definition of similarity may range, however, from very specific functionalities to very general concepts such as "bulkiness" of substituents. Other applications can be imagined, for example, in the field of organic synthesis. A comparison of the target molecule with possible precursors could be of valuable assistance in designing the best synthetic route.[4]

Determination of similar or common substructures has been largely a manual procedure. We feel, however, that this procedure is amenable to some degree of computer assistance. When large numbers of structures are involved or when the activity may be due to complex or disjoint substructures of varying size, manual examination cannot be expected to be exhaustive. In addition, automated procedures should be able to explore several different definitions of similarity in order to prevent particular biases from prejudicing the procedure.

It is perhaps a measure of the difficulty of this problem that only tentative steps have been taken to try to automate manual methods.[5,6] For some problems simple *topological* representations (i.e., simply the atoms of the molecule and their interconnections, devoid of any stereochemical information) of molecular structure are sufficient to characterize the structure prior to searching for common substructures. For example, fragmentation of molecules in a mass spectrometer is usually independent of stereochemical features, so that there is justification for relating the "activity" of fragmentation to molecular topology. In fact, published work by McLafferty and co-workers[6] on the problem of detecting common substructures in a set of molecules was inspired in part by a problem in mass spectrometry.

In subsequent sections we present an algorithm to solve the problem of determining common substructures in a set of structures, based on molecular topology. Recent work on computer-aided examination of structural similarities[5,6] deals efficiently with pairs of structures but has some limitations, specifically: (1) because only pairwise comparisons of structures are possible, the approach requires an amount of time which is a squared function of the number of structures, in particular for n structures, $n!/(2! * (n - 2)!)$; (2) the procedure[6] returns only "maximal" substructures (see below), an important measure of commonality in some, but not all, problems; and (3) the procedure has only limited control of the meaning of "similarity". For example, as a result of both (2) and (3), the program[6] without significant modification cannot recognize the presence of a common steroid nucleus substituted at C-3 and C-17 in the example (Figure 9) in reference 6, something a chemist would recognize immediately. In our approach, described below, we have attempted to alleviate these limitations to provide a more generally applicable program.

Other applications where determination of common substructures is important include those where *stereochemical* features of molecules are responsible for the observed activity. These applications must consider the set of structures displaying the activity as a set of three-dimensional entities. A computer search for common substructures must then be a search through three-dimensional representations of structures. To our knowledge, there are no algorithms or programs which perform this task; it is primarily a manual procedure. Some progress has been made in comparing molecular conformations automatically once the active sites of the set of structures have been determined by other techniques.[7]

I. METHOD

Our approach to the problem was to develop an algorithm which could be used as the basis for a computer program whose computation time would be linearly rather than exponentially, dependent on the number of structures. This would give some assurance that large numbers of structures could be examined for common substructures. In addition, the algorithm must be flexible enough to accommodate different chemical interpretations of similarity and allow a variety of descriptors of atom and bond properties (e.g., atom names, number of neighbors, hybridization) to be used to characterize structures and substructures. Our current facilities for representing and manipulating chemical structures for computer-assisted structure elucidation[8] deal primarily with topological representations of structure, with extensions to configurational stereoisomerism only a recent addition.[9] Thus, the algorithm described deals only with topology; in the conclusion we discuss ways of extending the approach to three-dimensional representations of structure. We begin the discussion of the method with a summary of several different ways in which similarities among structures may be described in topological terms.

Table I. Criteria for Expressing Similarities among Structures in the MAXSUB Program

Similarities Based on Descriptions of Substructures

(1) Cyclic and acyclic parts of structures:
  (a) atoms in rings and chains are equivalent; or
  (b) atoms in rings are not equivalent to atoms in chains, but ring size is irrelevant; or
  (c) same as (b), but atoms in rings differentiated by ring membership

(2) Atom type:
  (a) atoms differentiated by atom type; or
  (b) all atom types the same; or
  (c) chemist-defined equivalence of certain atoms or larger parts of structures

(3) Bond multiplicity:
  (a) bonds differentiated by their multiplicity; or
  (b) all bond multiplicities equivalent

Similarities Based on Attachment of Substructures within Structures

(1) "Simple"–method of attachment unimportant; or
(2) "Bonds"–number of bonds on each peripheral (see text) atom important; or
(3) "Hybridization"–multiplicity of bonds attaching peripheral atoms important; or
(4) "Exact"–combination of (2) and (3)
(5) "Pattern"–under the "simple" criterion (1), only the pattern of atoms is important.

A. DEFINITIONS OF SIMILARITY

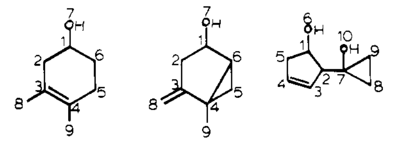
There are many possible ways to describe similarities among structures. In considering possible definitions of "common" substructures we have attempted to capture in the program those definitions which are most frequently used by chemists. In Table I we present the criteria available for selection by the chemist for expressing important similarities among structures. Each criterion is described in detail below.

Similarities Based on Descriptions of Substructures. There are three important criteria for similarity of structures based on descriptions of substructures which refer to characteristics of the atoms and bonds within the substructures themselves. The first is associated with properties of atom membership in rings or chains (which may be in part a function of the rest of the structure). The remaining two refer to types of atoms and bonds within substructures.

(*1*) *Cyclic and Acyclic Parts of Structures.* The chemist has the option of discriminating (or not) among atoms which are members of rings or chains of atoms. Three options are available (one must be selected):

(a) Atoms in cyclic and acyclic parts of structures are considered equivalent. Under this criterion, a chain of four atoms in a hexane molecule is the "same" as a connected substructure of four atoms in a cyclohexane molecule.

(b) Atoms in cyclic and acyclic parts of structures are considered nonequivalent, but atoms which are part of ring systems of different sizes are considered to be equivalent.



1 2 3

```
Similarity Criteria                          Largest Common Substructure(s)

I.   Ring/chain atoms: equivalent             X
     Bond multiplicities: not equivalent      .
     Atoms: equivalent                        X-X-X-X-X-X
     Attachment: simple
                                                  4
     Maximum size: 7
                                              1: 2 1 7 6 5 4 9
                                              2: 7 1 2 3 4 9 5
                                              3: 4 5 1 2 7 10 8
                                                 4 5 1 2 7 10 9

II.  Ring/chain atoms: equivalent                 C
     Bond multiplicities: equivalent              |
     Atoms: not equivalent                    C-C-C-C-C-C
     Attachment: simple
                                                  5
     Maximum size: 7
                                              1: 8 3 2 1 7 5 6
                                              2: 8 3 2 1 7 5 6
                                              3: 4 5 1 6 2 7 9
                                                 4 5 1 6 2 7 8

III. Ring/chain atoms: equivalent
     Bond multiplicities: not equivalent    C               C
     Atoms: not equivalent                  .
     Attachment: simple                   C-C-C-C-C       C=C-C-C-C        C-C-C-C-C-C

     Maximum size: 6                        6               7                  8

                                          1: 3 2 1 7 6 5  1: 4 3 2 1 7 6   1: 9 4 5 6 1 7
                                          2: 3 2 1 7 6 5  2: 8 3 2 1 7 6   2: 7 1 2 3 4 5
                                          3: 4 5 1 6 2 7  3: 4 3 2 7 10 8     7 1 2 3 4 9
                                                             4 3 2 7 10 9  3: 10 7 2 1 5 4

                                                             C
                                                             |
                                                         C-C-C-C-C

                                                             9

                                                         1: 4 5 1 6 7 2
                                                            8 3 2 1 7 6
                                                         2: 9 4 6 1 7 2
                                                         3: 9 7 2 1 6 5
                                                            8 7 2 1 6 5
                                                            5 1 2 7 10 8
                                                            5 1 2 7 10 9

Similarity Criteria                          Largest Common Substructure(s)

IV.  Ring/chain atoms: equivalent
     Bond multiplicities: not equivalent      C
     Atoms: not equivalent
     Attachment: bonds                        -C-C-

     Maximum size: 3                          10

                                              1: 7 1 6
                                                 7 1 2
                                              2: 7 1 2
                                              3: 6 1 5

V.   Ring/chain atoms:equivalent
     Bond multiplicities: not equivalent        C                C
     Atoms: not equivalent                      .                |
     Attachment: hybrid                       =C-C-C-C-C       C=C-C-C-C

     Maximum size: 6                            11               12

                                              1: 3 2 1 7 6 5   1: 4 3 2 1 7 6
                                              2: 3 2 1 7 6 5   2: 8 3 2 1 7 6
                                              3: 4 5 1 6 2 7   3: 4 3 2 7 10 8
                                                                  4 3 2 7 10 9

VI.  Ring/chain atoms: equivalent
     Bond multiplicities: not equivalent      C
     Atoms: not equivalent                    |
     Attachment: exact                        -C-C-

     Maximum size: 3                          13

                                              1: 7 1 2
                                                 7 1 6
                                              2: 7 1 2
                                              3: 6 1 5

VII. Ring/chain atoms: equivalent
     Bond multiplicities: not equivalent        C                C
     Atoms: not equivalent                      |                |
     Attachment: pattern                      C-C-C-C-C        C=C-C-C-C        C-C-C-C-C-C

     Maximum size: 6                            14               15                16

                                              1: 3 2 1 7 6 5   1: 4 3 2 1 7 6   1: 9 4 5 6 1 7
                                              2: 3 2 1 7 6 5   2: 8 3 2 1 7 6   2: 7 1 2 3 4 5
                                              3: 4 5 1 6 2 7   3: 4 3 2 7 10 8     7 1 2 3 4 9
                                                                  4 3 2 7 10 9     7 1 6 5 4 3
                                                                  3 4 5 1 6 2   3: 10 7 2 1 5 4
                                                                                   6 1 2 7 8 9

                                                   C
                                                                                 
                                              C-C-C-C-C        C=C-C-C-C-C

                                                  17               18

                                              1: 4 5 1 6 7 2   1: 4 3 2 1 6 5
                                                 8 3 2 1 7 6   2: 8 3 2 1 6 5
                                              2: 9 4 6 1 7 2   3: 4 3 2 7 8 9
                                              3: 9 7 2 1 6 5
                                                 8 7 2 1 6 5
                                                 5 1 2 7 10 8
                                                 5 1 2 7 10 9
```

Figure 1. Largest common substructures, under different measures of similarity, among structures **1–3**. The lists of numbers associated with each common subgraph correspond to the atom numbering given for structures **1–3**. These lists are written in an order which facilitates detection of the common substructures within each structure. No correspondence of numbers relative to the drawn substructures is intended.

Under this option, a carbon atom of cyclohexane is considered to be equivalent to a carbon atom in cyclopentane, but not equivalent to a carbon atom in hexane.

(c) Atoms in a ring are considered nonequivalent to atoms in a chain and atoms in rings are further differentiated according to ring membership by associating with these atoms a value which is the size of the smallest ring to which the atom belongs. For example, C-4 in structure **2** (Figure 1) which is in a six-, five- and three-membered ring system will be assigned the value of three, the smallest ring. Only atoms which have the same ring value will be considered to be equivalent.

The ring properties of atoms, if option (b) or (c) is selected, are used in a preprocessing step of the algorithm (below) to associate this property with every atom in every structure. The algorithm then proceeds with atoms whose identities are based in part on the property of ring size. If option (a) is selected, the structures are not affected.

(*2*) *Atom Type, Functional Groups, and Larger Parts of Structures*. There is some flexibility provided for defining certain parts of structures to be the "same" based on type of atom or functionality:

(a) Atoms of different type are considered to be different. This option would be selected normally,

(b) All atom types are considered to be equivalent. If the chemist is not concerned about the type of atom, i.e., wants to refer to C, N, O, etc., as the "same", MAXSUB automatically, in a preprocessing step, changes all atom names in all structures to type "any". The algorithm proceeds allowing any atom type to be equivalent to any other. This is illustrated in Figure 1. When all atoms are considered to be equivalent MAXSUB determines the largest common substructure in structures **1-3** to be **4**, containing seven atoms. C-9 in structure **2** is the "same" as 0-7 in **1** and 0-10 in **3**.

(c) If more specific statements about equivalence of atoms, functional groups, or larger parts of the structures are required, a different preprocessing of the structures is required. Using our structure manipulation program REACT,[7] the chemist can define the required equivalencies (REACT and MAXSUB can communicate via files of structures). For example, different halogen substituents can be converted by REACT to one common substituent named Hal, and subsequently be treated by the program as equivalent atoms. Different functionalities can be made equivalent for MAXSUB analysis, in the same way. The method is quite general in that *any* feature of the set of structures which the chemist wants to consider as common can be converted by REACT into a single "atom" prior to analysis by MAXSUB. This is illustrated for structures sharing common portions of a steroid nucleus in a later example.

(*3*) *Bond Multiplicity*. Two options are available which concern multiplicity of bonds within the substructure:

(a) Bonds are differentiated by their multiplicity. A single bond is never the "same" as a double bond, and so forth.

(b) In instances where only the overall connectivity of the substructure is important but not the exact multiplicity of the bonds, the chemist can ask the computer to compare his structures with the requirement that all bond multiplicities be considered equivalent.

If the bond multiplicities are selected to be equivalent, all multiple bonds are removed from all structures in a preprocessing step (below). The algorithm then proceeds with the modified structural representations.

Similarities of Attachment of Substructures within Structures. We can regard a substructure as a structural fragment which was cut out of a larger structure. Such a substructure will contain two type of atoms between which we differentiate in MAXSUB those atoms "internal" to the substructure and those atoms "peripheral" to the substructure. The internal atoms are those which have bonds *only* to other atoms in the substructure and are therefore completely contained within the substructure. Peripheral atoms have one or more "external" bonds which represent points of attachment of the atoms of the substructure within the larger structure. As an illustration, consider the substructure **6** which is common to structures **1–3** under the criteria given in Figure 1.III. By our definition, C-3 and C-5 are peripheral atoms in structure **1** and represent the atoms used to attach the substructure **6** within **1**. In structure **2**, C-3, -5, and -6 are peripheral atoms because they are involved in attaching **6** within **2**. Similarly, C-1, -2, and -6 and O-7 in **1** and C-1 and -2 and O-7 in **2** are internal atoms.

We make the distinction between peripheral and internal because the chemist's view of similarity among structures frequently includes specification of the precise method of attachment of a given substructure with each member of the set of structures. We provide in MAXSUB the capability to specify how peripheral atoms are connected to the remainder of a structure. The modes of operation of MAXSUB to express these connections are as follows.

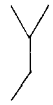
(*1*) *Simple.* In this mode the method of attachment of the peripheral atoms to atoms external to the substructure is not important. Thus, the two-atom substructures CH_3-CH_0, CH_3–CH_1, and CH_3–CH_2 are considered equivalent under this criterion of similarity. In fact, there is *no* distinction between periphral and internal atoms in this mode. This is illustrated further in Figure 1.I–1.III. In each of these examples of largest common substructures the attachment characteristics were designated as "simple". Thus, for example, in Figure 1.I, the internal atom C-1 in structure **1** is the "same" as the peripheral atom C-4 in **2** (both represent the branch point of the common substructure **4**).

(*2*) *Bonds.* In this mode the number of bonds connecting peripheral atoms to the larger structure is important. Bonds are counted *only* to nonhydrogen, neighboring atoms and double bonds represent two bonds, triple bonds three, and so forth. In other words, this criterion is related directly to the number of hydrogen atoms on each peripheral atom. Thus, CH_3–$CH<$ is the "same" as CH_3–$CH=$ but not CH_3–CH_2– or CH_3–CH_0. In the example (Figure 1.IV), the largest common substructure, of structures **1**–**3**, based on this definition of similarity, is of size three, a hydroxyl group connected to a CH connected in turn to a CH_2 (**10**). These structures share no larger common substructure under this definition.

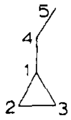
(*3*) *Hybridization.* Under this criterion, the hybridization of each peripheral atom is important. Thus, CH_3–CH_1= will match CH_3–CH_0= but not CH_3–$CH_1<$. Since the number of hydrogens attached to peripheral nodes is not important under this criterion, structures **1–3** have two common substructures of size six, **11** and **12** (Figure 1.V), but none larger.

(*4*) *Exact.* This mode is the combination of (2) bonds and (3) hybridization. Peripheral atoms are the "same" only if their hybridization and number of attached hydrogen atoms are the same. Thus, CH_3–$CH=$ will only be the same as CH_3–$CH=$. In the example (Figure 1.VI), the largest common substructure is the same as that obtained in Figure 1.IV. Generally, the added requirement of equivalence of both bonds and hybridization results in smaller common substructures.

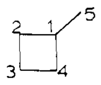
(*5*) *Pattern.* Structures **19-21** have no common subgraph of size five according to the previous definitions of structural similarities. However, every chemist will recognize that these three structures can be constructed from an isoprene skeleton **19**. All three possess the same "pattern" of atoms but differ in the ways in which rings are formed by bonds between atoms in the substructure. We can find an isoprene substructure in **20** by removing the bond between C-2 and C-3, or in **21** by removing the bonds between C-2,3 or C-3,4. We define this additional type of structural similarity as pattern similarity. Applying this definition of structural similarity to the example of Figure 1, the largest size of common pattern in structures **1–3** is six (Figure 1.VII). Four of the common patterns **14–17** appear also in the simple mode (Figure 1.III), but substructures **15** and **16** are found in additional places in the original structures. Substructure **18** is new and can be obtained only by using the pattern mode to express structural similarities.



19



20



21

B. PROGRAM OPERATION AND REPORTING RESULTS

Modes of Operation. There are three modes in which MAXSUB can be operated: automatic, manual, and selective. In the automatic mode the program will find *all* common substructures present in the set of structures, based on the previous definitions of structural similarity. In the manual mode the program will stop after finding common substructures of a size specified by the chemist. If no such substructures exist the program will stop after reporting the largest common substructure of any smaller size. In the selective mode the chemist can specify the order in which he wishes the structures to be examined. The program, after finding substructures common to all structures, will continue to report common substructures of larger size for a decreasingly smaller subset of the initial structure list, this subset including the first structure in the list. The program stops when it fails to find a substructure common to the first structure and at least one other structure in the structure list. This is particularly useful when the structural problem involves one particular molecular structure (which can be placed first on the list) and its relationship to a set of other structures.

C. REPORTING RESULTS

Heteroatoms. In any mode of operation, the program can be instructed to search for and report only those substructures containing at least one heteroatom.

There are three options to display results of the search for structural similarities:

(*a*) *All.* The program reports *all* common substructures of *all* sizes.

(*b*) *Maximal.* The program reports a common substructure *only* if this substructure is not *completely* included in a larger common substructure. This option is closely related to the definition of maximal substructure used by McLafferty and co-workers[6] generalized to a set of structures numbering more than two members.

(*c*) *Largest.* The program reports *only* the largest common substructure.

D. THE ALGORITHM

The problem of finding common substructures in a set of structures is simple to visualize. In order for a specific substructure to be common to the set of structures, every structure must possess at least one such specific substructure. Depending on the chemical problem, interest may be focused on all common substructures, substructures of a given size or range of sizes, "maximal" substructures (see above), or greatest common substructure. The precise meaning of common may be different in different chemical contexts and depends also on definition of what constitutes criteria for similarity (above). It is a challenge to devise a single algorithm which can yield all (or selected ones) of these possibilities and still retain sufficient efficiency to examine many structures.

Our approach to solving the problem resembles to some extent the way in which a chemist would examine a group of structures for similar features. The algorithm begins by searching for two-atom, common substructures and continues exploring, stepwise, ever larger substructures until appropriate terminating conditions are met (see below). The algorithm is essentially brute force, but admits of enough computational heuristics and simplifying conditions to increase dramatically its efficiency. The algorithm can even utilize available information, obtained from the chemist, about the presence of easily recognized common substructures, further simplifying the search.

Nomenclature. Since the algorithm deals with a graph theoretical problem, we shall use terms taken from graph theory. The topological representation of a chemical structure is a LABELED GRAPH, in which the atoms are NODES and the chemical bonds are the EDGES. A subset of connected nodes in the graph is a SUBGRAPH. (There is a one-to-one correspondence of graph to structure, subgraph to substructure, node to atom, and edge to bond in this topological representation of structure.) Nodes in the graph are labeled with numbers, NODE NUMBERS, and an edge is described by node numbers of the pair of nodes which it connects. The numerical representation of the graph is a CONNECTION TABLE. A connection table is constructed from CONNECTION TABLE ENTRIES. A connection table entry contains a NODE NUMBER, NODE TYPE, and a list of NEIGHBORS which is a list of the node numbers of the neighboring nodes to which the entry is connected.

ISOMORPHIC graphs are graphs (or subgraphs) which are equivalent but may have different connection table representations. For each such set of isomorphic graphs, we select one representative according to a previously described algorithm.[10] This representative is the CANONICAL representative of the graph, and the selection process is called CANONICALIZATION. Graphs which have the same canonical representation (i.e., are isomorphic) are said to be of the same ISOMORPHIC TYPE. The canonicalization procedure also computes a number which is related to the canonical representation of the graph. This number is called the CANONICAL KEY. Two graphs which have different keys are guaranteed to be nonisomorphic. Two graphs with the same key may or may not be isomorphic, and possible isomorphism must be checked by direct comparison of the canonical representations.

The following steps summarize our algorithm:

(1) Preprocess and Order the Structure List. Depending on the chemist's definition of similarity of substructures, various preprocessing steps are possible (see section I.A). The preprocessing results in a new characterization (representation) of the structures which makes subsequent analysis more efficient. Preprocessing is carried out, for example, when bond orders or atom types are deemed irrelevant, or when ring membership of atoms is deemed important.

The set of structures (the "structure list") to be analyzed is then reordered (in the selective mode of operation, the chemist-specified ordering of the structures is preserved). Since the final list of common subgraphs of size *m* is a subset of all the subgraphs of size *m* which can be obtained from any graph, it is important to start first with the graph which has the smallest number of possible subgraphs. There are fewer subgraphs in graphs which have smaller numbers of nodes of high degree (in chemical terms, smaller numbers of atoms which are highly substituted). There also tend to be fewer isomorphic types in such graphs. Therefore, the structure list is ordered by increasing number of atoms of high degree per structure. Note that this also tends to place smaller structures near the beginning of the list when diverse structures are being analyzed. This ordering reduces the number of subgraphs one must consider in the first step and, thus, in all subsequent steps.

(2) Generation of Common Subgraphs of Size Two. (*a*) *Construction of the Initial Subgraph List.* A list is created of all connected, two-atom subgraphs which are present in the *first* structure in the ordered structure list. The number of these subgraphs is equivalent to the number of edges in the graph representing the first structure. The trivial case of single atom subgraphs is ignored.

(*b*) *Labeling the Nodes.* Based on the chemist's definition of structural similarity, the peripheral nodes in the subgraphs are labeled according to their method of attachment in the original graph. The labeling procedure associates with the nodes properties related to the chemist's criterion of similarity of attachment described in the previous section (for example, number of bonds, or hybridization). The labeling process is an independent routine and with minor modification can accommodate additional node descriptors such as stereochemistry or other chemical or physical properties of atoms.

(*c*) *Canonicalization.* The labeled subgraphs are transformed to their respective canonical representations.

(*d*) *Sorting and Grouping the Subgraphs into Nuclei.* The canonical representations of the subgraphs are sorted in order of increasing canonical keys, and subgraphs with the same key are checked to see if they are of the same isomorphic type. Each unique, canonical representation is termed a NUCLEUS, to convey the sense of growth, or expansion, of the nuclei in subsequent steps. A record is maintained for each nucleus which relates the nodes of the nucleus to nodes of the graph for every place in the graph where the subgraph was found (obviously one may encounter the same subgraph in more than one location in the structure). For efficiency, only one copy of a nucleus of the same isomorphic type is kept if the same subgraph is encountered in subsequent structures; only the record relating each nucleus with its origins is updated. The sorted nuclei list becomes the MASTER LIST which records all common or candidates for common subgraphs to this point.

(*e*) *Examining the Remaining Structures.* Steps a–d are repeated for the next structure (the *second* structure the first time through this loop). Each nucleus which is generated in step d is compared to the nuclei already on the master list from previous steps. There are three possible results of the comparison, as follows: (1) if the nucleus is found to be equivalent to a nucleus in the master list, only the original nucleus in the master list is kept but now together with a record of the origin of the nucleus in the structure under examination (this nucleus remains as a candidate for a common substructure); (2) if the nucleus is not equivalent to any of the nuclei on the master list, then the subgraph represented by this nucleus *cannot* be common to all the structures (this new nucleus is *not* added to the master list); or (3) nuclei on the master list are not encountered in the structure. After comparison of all the nuclei from the new structure to the master list, the list is searched for nuclei which were not encountered in the new structure. These nuclei are deleted from the master list because if they were not encountered in the new structure they cannot be common to the set of structures.

This procedure is repeated for the next structure in the structure list and continues until either all the nuclei are deleted from the master list or all the structures in the structure list are examined. In the first case, in the manual or automatic mode of operation, the procedure is terminated because there are no common subgraphs of size two. In the selective mode, structures which do not contain any of the nuclei on the master list are marked so that they are never considered in subsequent steps of the procedure and the procedure is continued for the remaining structures. If all structures have been examined, the algorithm proceeds to the next step (f).

(*f*) *Marking Extraneous Edges.* After all structures have been considered for subgraphs of a given size, size two for the first iteration, every structure is examined for edges which no longer need be considered as candidates for participation in common subgraphs. Based on the information in the master list, the program marks *all* edges in every structure in the structure list which *do not participate* in any of the common subgraphs. It can be shown that an edge which is not involved in any of the common subgraphs to this point *cannot* be a part of any larger common subgraph. Marking them prior to step 3, below, prevents them from being considered in any subsequent step of searching for larger common subgraphs.

(3) Growing the Nuclei. When all structures have at least one common subgraph of size *m*, beginning with size two, the program continues by checking for the presence of common subgraphs of size *m* + 1. At this stage the master list contains all common nuclei of size *m*, and with each nucleus the associated information relating each nucleus to each occurrence of the subgraph in each structure. Beginning again with the first structure in the list, for each occurrence of each nucleus the program creates new subgraphs of size *m* + 1 by adding *one* neighbor node in all possible ways. Neighbor nodes are added only if they are connected to the subgraph by unmarked edges (see step 2.f). For the first structure, all nuclei created in this way are added to the master list. Each subsequent structure is considered in turn by growing nuclei and looping through steps 2.b–e. When all structures have been checked for nuclei of size *m* + 1, step 2.f is repeated to mark extraneous edges. The growing procedure is continued as long as the master list contains common nuclei.

(4) Generation of Patterns. Patterns (see section I.A) are generated from the subgraphs (step 3) by removing edges connecting peripheral atoms to other atoms in the subgraph. Each subgraph obtained this way is a candidate pattern and is treated by steps 2.c–f of the algorithm. An edge to a peripheral node will not be removed if removal creates a disconnected subgraph.

Efficiencies in the algorithm are achieved in four important ways: (1) the initial ordering of the structure list tends to reduce the computational effort; (2) the ordering of the subgraphs according to their isomorphic type eliminates the problem of comparing subgraphs which are obviously different by their canonical keys; (3) commonality of subgraphs is checked as each structure is processed, thereby removing subgraphs immediately from consideration which are not in common; and (4) the method of marking edges in structures prevents having to consider these edges in subsequent steps as larger subgraphs are grown. In many cases, given some degree of structural diversity in the list of structures analyzed, most edges are marked early in the procedure, and the number of neighbors which can be selected to grow larger subgraphs diminishes rapidly.

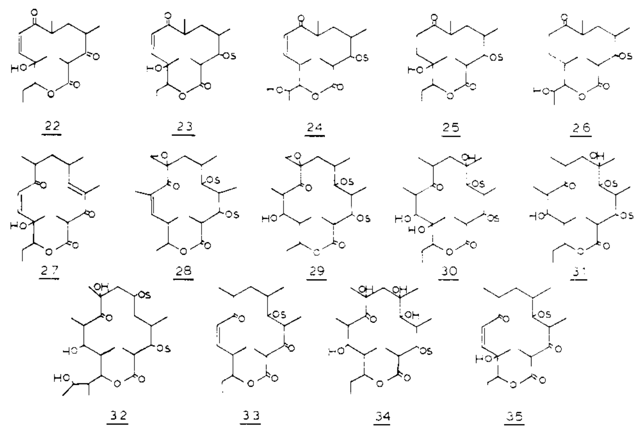


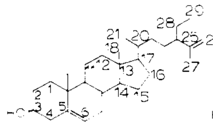
Figure 2. Selected macrolide antibiotics **22–35** studied by MAXSUB (S represents the sugar moiety).

II. EXAMPLES OF APPLICATION

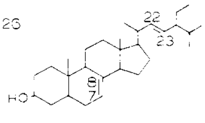
We present in this section two brief examples to illustrate different types of problems for which the MAXSUB program might be useful. In addition, we illustrate the different results obtained for different specifications of similarity among structures.

Because of our past interest in the field of macrolide antibiotics, we examined selected macrolides, possessing 12- and 14-member rings, for structural similarities. These are relatively large molecules containing between 22 and 33 nonhydrogen atoms. For our test we selected five 12-member and nine 14-member macrolides, **22–35** (Figure 2). The results of the search for structural similarities are summarized in Figure 3.

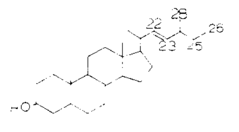
When the definition of similarity of attachment is "exact", there are only two substructures common (maximal) to all 14 structures. One is of size two (**36**) and contains a tertiary and secondary carbon. The second (**37**), of size six, describes the environment of the common lactone functionality, a tertiary carbon attached to the alcoholic oxygen of the lactone and a methyl group which is connected through a tertiary carbon to the carbonyl of the lactone. Loosening the definition of structural similarity, and allowing atoms with different substitution of hydrogens to match, increases the number and the size of the common subgraphs. For example, in the "simple" mode of attachment and when atoms in different ring sizes are considered to be equivalent, but different from atoms in chains, there are three different sizes of common (maximal) substructures **38–43** (Figure 3). The largest common substructure, **43**, is of size 13 and describes a more extended environment around the common lactone functionality.



44



45



46

As another illustration, recent investigations of the mass spectral fragmentations[11] of marine sterols revealed several similar fragmentation pathways. We have used MAXSUB to explore selected sterols for common substructures which might be related to fragmentation. Structures **44–46** are three such compounds selected for study. These compounds obviously share a common steroid nucleus, a 3-hydroxyl substituent and a portion of the side chain. In fact, if MAXSUB is operated ignoring bond multiplicities and requiring exact

```
Program parameters


Mode of operation: automatic
Ring/chain atoms: equivalent
Bond multiplicities: not equivalent
Atoms: not equivalent
Attachment: exact
Displaying results: maximal subgraphs


                    Common subgraphs of size 2

                              |
                            -C-C-
                             36


                    Common subgraphs of size 6

                               O
                          |    ||
                        -C-O-C-C-C
                                 |

                             37

----------------------------------------------------------------


Program parameters


Mode of operation: automatic
Ring/chain atoms: not equivalent
   All ring atoms are equivalent
Bond multiplicities: not equivalent
Atoms: not equivalent
Attachment: simple
Displaying results: unique


                    Common subgraphs of size 5^a

                       O  C            O  C
                       |  |            || |
                    C*-C*-C*        C*-C*-C*
                       38              39


                    Common subgraphs of size 8

    O           C             C           C       O  C     C
    ||          |             |           |       || |     |
 C*-C*-C*-C*-C*-C*         C*-C*-C*-C*-C*-C*     C*-C*-C*-C*-C*

      40                         41                    42


                    Common subgraphs of size 13

                        C  C     O  C     C
                        |  |     || |     |
                     C*-C*-C*-O*-C*-C*-C*-C*

                             43


(a) * refers to atoms in the ring system.
----------------------------------------------------------------
```

Figure 3. Common substructures found by MAXSUB for macrolide antibiotics **22–35** under two different measures of similarity.

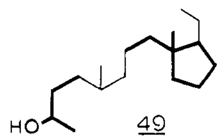
attachment of peripheral atoms, a common substructure similar to **46** (without the double bond and C-25, C-26 and C-28) is found as the largest common subgraph, as expected.

Of more importance are examples of results obtained when MAXSUB is instructed to consider chemist-selected substructures to be in common from the beginning. As an illustration, two substructures were selected as common to the three structures, one comprising C-1–C-4 with the hydroxyl substituent, the other C-11–C-18 (both indicated in bold face in the substructure **49**, Figure 4). Using the REACT program, these two substructures were converted into single "atoms" and the modified representations of **44–46** were analyzed by MAXSUB. In this instance, bond multiplicities were considered important (see criteria for similarity summarized in Figure 4). Results of the search for common substructures are summarized in Figure 4. Note that in specifying an exact mode of attachment, substructure **49** is the largest common substructure; the double bonds at C-5,6 and C-25,26 in **44**, C-7,8 and C-22,23 in **45** and C-22,23 in **46** prevent further "growth" to larger substructures under this criterion.[12]

```
Program parameters                       Common subgraphs of size 3

Mode of operation: automatic
Ring/chain atoms: equivalent                  | |         |
Bond multiplicities: not equivalent         -C-C-C-     -C-C-C
Atoms: not equivalent
Attachment: exact                            47          48
Displaying results: unique

                                         Common subgraphs of size 18
```



49

Figure 4. Common substructures found by MAXSUB for marine sterols **44–46**.

III. CONCLUSION

In designing an algorithm for finding structural similarities, we followed an approach we feel is related to some extent to that taken by chemists. Start by exploring similarities based on small substructures. If similarities are found, continue by exploring ever larger substructures until a point is reached beyond which the set of structures appears to have nothing in common. This approach is more efficient than the technique of pairwise comparison when dealing with large numbers of structures. Also, MAXSUB solves the problem of finding common subgraphs according to a wide variety of definitions of structural similarity. The algorithm is especially efficient in the examination of variety of structures when operating in the most restrictive mode of definition of structural similarity. The total central processor (cpu)[13] time used for generating and displaying the results for the 14 macrolides in the exact mode was 213 seconds. However, operating in the more general simple mode, the program used a total of three hours cpu time. This dramatic increase in time results because all peripheral and internal atoms which are, for example, carbon atoms, are considered the "same" and are included in small substructures. Thus, few edges can be marked in early steps of the procedure and large numbers of new substructures must be grown at every step.

Utilizing knowledge about the presence of common substructures reduces considerably the amount of computation time and allows the chemist to focus on nontrivial results. Such information reduced the amount of computation time for finding maximal subgraphs for the three marine sterols (Figure 4) from several minutes to 45 seconds.

These figures of cpu time are somewhat misleading because one portion of the program is in the language INTERLISP which tends to be inefficient. The time-consuming step in our algorithm is the step of canonicalization. Although the canonicalizer is written in a more efficient language (SAIL, an Algol variant), the overhead of communication between INTERLISP and SAIL portions of the program contributes to inefficiency. If the algorithm were entirely in a more efficient language, we would expect a factor of 10 to 100 decrease in running time, a factor based on our experience in converting similar programs to more efficient languages. We are investigating an exportable version of MAXSUB which would have this desired efficiency.

Another way to reduce the computation time is to find a more efficient way to canonicalize structures or reduce the number of necessary canonicalization steps. The former is unlikely to yield much improvement because of the effort that went into studies of efficiency to begin with. The latter might yield to a more intelligent procedure for growing nuclei. The part of the algorithm which grows nuclei adds one atom at time. This procedure can generate many duplicate subgraphs by following different paths. Prospective elimination of such duplicates is desirable and we are working on this problem.

Since our structural representation is topological, information about the geometry of the resulting common subgraphs is not available. We consider this a serious limitation for studying biological structure activity relationships. Work on entering stereochemical information is in progress.[9] As was mentioned in the discussion, associating properties with atoms and bonds is done in one module of the program and even now the set of descriptors could be enlarged in ways which might implicitly include some geometrical information, e.g., as ring membership does already. Addition of other structural properties, such as stereochemistry of certain atoms and relative locations of atoms in a structure, is possible but more difficult to implement in a general way.

REFERENCES AND NOTES

(1) Part XXX of the series "Applications of Artificial Intelligence for Chemical Inference". For part XXIX see J. G. Nourse, R. E. Carhart, D. H. Smith, and C. Djerassi, *J. Am. Chem. Soc.*, **101**, 1216 (1979).

(2) We wish to thank the National Institutes of Health (RR-00612, GM20832, and GMO6840) for their generous financial support for our research and for their support (RR-00785 SUMEX) of the SUMEX computer resource on which the MAXSUB program is available via nationwide computer networks.

(3) A. J. Stuper and P. C. Jurs, *J. Am. Chem. Soc.*, **97**, 182 (1975).

(4) E. J. Corey, W. T. Wipke, R. D. Cramer, III, and W. J. Howe, *J. Am. Chem. Soc.*, **94**, 421, 431 (1972).

(5) (a) J. E. Armitage and M. F. Lynch, *J. Chem. Soc. C*, 521 (1967); (b) J. E. Armitage, J. E. Crowe, P. N. Evans, M. F. Lynch, and J. A. McGuirk, *J. Chem. Doc.*, **7**, 209 (1967); (c) J. M. Harrison and M. F. Lynch, *J. Chem. Soc. C*, 2082 (1970).

(6) M. M. Cone, R. Venkataraghavan, and F. W. McLafferty, *J. Am. Chem. Soc.*, **99**, 7668 (1977).

(7) F. A. Gorin and G. R. Marshall, *Proc. Natl. Acad. Sci. U.S.A.*, **74**, 5179 (1977).

(8) (a) T. H. Varkony, R. E. Carhart, and D. H. Smith in "Computer Assisted Organic Synthesis", W. T. Wipke and W. J. Howe, Eds., American Chemical Society, Washington, D.C., 1977, p 188; (b) R. E. Carhart, T. H. Varkony, and D. H. Smith in "Computer Assisted Structure Elucidation", D. H. Smith, Ed., American Chemical Society, Washington, D.C., 1977, p 126; (c) T. H. Varkony, R. E. Carhart, D. H. Smith, and C. Djerassi, *J. Chem. Inf. Comput. Sci.*, **18**, 168 (1978).

(9) (a) J. G. Nourse, R. E. Carhart, D. H. Smith, and C. Djerassi, ref 1; (b) J. G. Nourse, *ibid.*, **101**, 1210 (1979).

(10) R. E. Carhart, MIP-R-118, Machine Intelligence Research Unit, University of Edinburgh, United Kingdom, 1977.

(11) C. Djerassi, *Pure Appl. Chem.*, **50**, 171 (1978).

(12) Using this method, chemist-selected portions of structures are converted by REACT to single "atoms", thereby losing information on the relative points of attachments with the rest of the structure. If the portion so replaced has points of attachments which are *symmetrically equivalent*, then there is no problem in interpretation of results. Otherwise, there is ambiguity in selecting points of attachment which in some instances may yield substructures which are not in fact in common.

(13) The program runs on a Digital Equipment Corp. PDP-10 computer at SUMEX computer facility, Stanford University. The program is available (to the limit of available resources) on-line to interested persons over three nationwide computer networks. For information on access to the program please contact the authors.

A Representation of π Systems for Efficient Computer Manipulation

JOHANN GASTEIGER

Institut für Organische Chemie, Technische Universität München, D-8046 Garching, West Germany

Received September 18, 1978

A data structure has been developed which is particularly well suited for the representation of π systems. This representation is based on a separation of σ, π, and n electrons. All valence electrons in molecules and reactions can be accounted for, giving algebraic properties to this description. The representation is excellently amenable to computer manipulation, and programs based on it have been developed. It can also serve as an interface to valence bond or molecular orbital calculations.

The automatic manipulation of chemical structures has become of increasing importance. Documentation and information retrieval, as well as deductive computer programs, ask for efficient data structures.[1] In particular, the selection of a representation for chemical compounds can greatly influence the performance of a system. For certain problems special data structures may offer advantages, but for a general-purpose system an unambiguous topological representation seems indispensable. Fragmentation codes and linear codes have previously been chosen because of their compactness, but for certain problems they might need decompressing and data conversion. Furthermore, some primary information on the topology of a structure might no longer be accessible.

With the advances in computer technology, storage space has lost its crucial importance. Rather, it is more desirable that in the design of a system such a representation for molecular structures is chosen which retains direct information on atoms and bonds. This provides for an open-ended approach to the manipulation of molecules. Thus, in computer programs for chemical problems, topological representations have gained a similar importance as has the structural formula for conventional chemical communication.

In a topological representation for molecular structures information on the atoms of a molecule and their connectivities is provided. In the program systems that we are developing[2-4] this information is enriched by data on bond orders and free electrons. Thus, in effect, account is taken of all valence electrons of a molecule. This allows immediate access to valence states, charges, etc. It also provides the basis for the exhaustive generation of chemical reactions which are considered as bond and electron shifting processes and are generated accordingly.[2-4]

This work was initiated by a mathematical model of constitutional chemistry.[5,6] In this model molecules and ensembles of molecules are represented by so called BE