ronmental fate related information contained in the data bases were successfully alphabetized. In the case of hyphenated trade names, the hyphen was omitted from the name to avoid problems.

One advantage of the algorithm is that one can identify problem names by inspection. Many classes of compounds and forms of nomenclature are easily accommodated by an extension or modification of the algorithm. For example, when chemical names are written in inverted order (e.g., "acetic acid, chloro-", or "2-propenoic acid, 3-phenyl, 2-methylpropyl ester") one must modify the algorithm as follows: eliminate final hyphens or hyphens preceding commas from the sortkey, replace all commas with blanks, and then apply the standard algorithm. This modified version of SORTKEY could be used with noninverted as well as inverted nomenclature. Another situation where the algorithm can be modified to handle the chemical names is with bridged and fused ring systems. In these cases locants may appear within square bracked (e.g., "1,4-diazobicyclo[2.2.2.]octane" and "indeno[1.2.3-*cd*]-pyrene"). For these chemical names, the locant contained within the square brackets should be eliminated from the sortkey before the standard algorithm is applied.

SORTKEY may fail in situations where names of stereoisomers and optical isomers are to be sorted because there are situations where two locants separated by a hyphen are adjacent (e.g., *cis*-2-butene). These situations can be easily identified. Sometimes the chemical name can be rewritten so the algorithm applies (e.g., "2-butene (cis)" or "2-butene, cis-"). Alternatively, one can search for certain character strings (e.g., cis, dl) which remain at the beginning of the word after the algorithm is applied and remove these locants from the primary sortkey. One could further extend the algorithm in these cases by specifically searching for specific character strings that arise with stereoisomers or optical isomers throughout the chemical name, but this is beyond the intent of this paper. We have also avoided dealing with chemical names that have subscripts or superscripts because there is no simple way of entering them into a computer at a terminal without some special designation. The primary strengths of using the SORTKEY algorithm for alphabetizing computer files containing chemical names is its simplicity and the fact that its implmentation is possible for most people. The version presented is sufficient for many situations, and several modifications to the algorithm can be made to extend its applicability.

ACKNOWLEDGMENT

The support for this work from the U.S. Environmental Protection Agency, Office of Pesticides and Toxic Substances, under Cooperative Agreement CR 806902020 is gratefully acknowledged.

REFERENCES AND NOTES

(1) Howard, Philip, H.; Sage, Gloria, W.; LaMacchia, A.; Colb, Andrew. "The Development of an Environmental Fate Data Base". *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 38-44.
(2) International Union of Pure and Applied Chemistry. Organic Chemistry Division. Commission of the Nonmenclature of Organic Chemistry. "Nomenclature of Organic Chemistry, Sections A, B, C, D, E, F, and H"; Pergamon Press: Oxford, 1979.
(3) "Chemical Abstracts Index Guide"; Chemical Abstracts Service: Columbus, OH, 1982.

Computer-Assisted Examination of Compounds for Common Three-Dimensional Substructures[1]

CHRISTOPHER W. CRANDELL*† and DENNIS H. SMITH‡

Department of Chemistry, Stanford University, Stanford, California 94305, and Lederle Laboratories, Pearl River, New York 10965

Received April 28, 1983

A program for finding common three-dimensional substructures within a set of chemical compounds is described. The program allows a user to define what constitutes commonality of substructures by providing control over the importance of degree of substitution, atom type, aromaticity, and hybridization. Simple examples are used to illustrate various phases of the search process, and an application of the program to a structure/activity problem is used as a more realistic example.

(1) INTRODUCTION

Comparison of structural features within a set of chemical compounds that display a common property is a frequent problem in chemical research. This problem can usually be characterized as one of relating the structure of the compounds 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 compounds all display a characteristic pattern or subpattern in a spectroscopic technique or biological in that the compounds demonstrate similar physiological effects.

One generally makes the assumption that the common activity of a set of compounds is due to some similar structural feature or features. Establishing the relationship between common features and activities is of obvious importance. In biological applications, these relationships are useful, for example, in designing new drugs. In spectroscopic applications, these relationships provide data, for example, for correlation tables. The definition of structural similarity and what constitutes a structural feature are, of course, dependent on the specific application, but the problem can be stated and solved for the general case.

This paper describes an algorithm, currently implemented in an interactive computer program, for comparing three-dimensional (3-D) representations of structures to find 3-D common substructures that represent hypothetical requirements for activity, of a general nature. Such comparisons are generally done manually by using molecular models. There are obvious limitations to manual methods, particularly when

† Stanford University.
‡ Lederle Laboratories.

0095-2338/83/1623-0186$01.50/0

two or more structures must be superimposed or accurate measurement of distances is required. Over the past few years, a variety of computational procedures have been applied to the problem of structural comparison. Although none of these procedures directly addresses the problem solved by our method, a brief review will help place our method in the context of related approaches.

Two computer programs have been described that compare graphical, or topological, representations of structures for common substructures. One program[2] compares structures in a pairwise fashion to identify commonalities. Another program[3] compares a set of structures for common substructures. Both programs search for *connected* substructures. Neither program considers any geometrical aspects of structures. These are obvious limitations for structure/activity relationships that depend on interactions of remote intrastructure groups and/or stereochemical features. An advantage of these programs is that they can function without initial assumptions about what structural features might be in common.

Several approaches to comparison of 3-D representations of structures for common features have been described. One approach is 3-D substructure search,[4,5] which determines the presence or absence of a designated 3-D substructure in a set of structures. This approach does not solve the problem of identifying important 3-D substructures; it can only test for their presence once they are proposed. An extension of one method[5] to the problem of identifying common substructures has been proposed but not demonstrated. In our opinion, the proposed algorithm[5] would be too inefficient to be of practical utility because time consuming computations would be performed in x,y,z coordinate space.

For comparison of complete 3-D structures, computer graphics can be used to superimpose two or more structures. This aids in visualizing gross comparisons. Quantifying the degree of overlap, however, requires a computer program; several approaches have been suggested.[6-8] Some techniques[6,7] are limited in the total number of structures that can be compared. More importantly, the approaches all depend on initial assumptions about common structural features to provide anchor points in coordinate space around which the rest of the structures of conformations are placed. In other words, a hypothesis is made about common portions of the structures responsible for activity, and then structures are compared for their ability to meet the geometrical requirements.

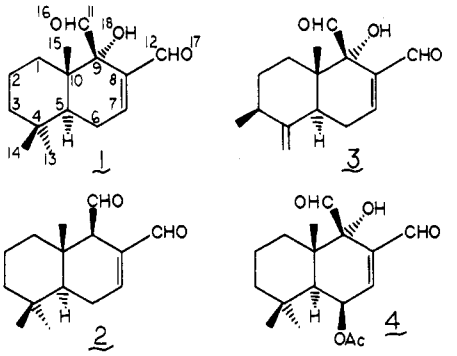
A recent report has proposed a method for automated selection of pharmacophores, including those based on geometric representations of structures.[9] This approach uses a distance matrix representation of structures similar to our representation (see below). This approach has several limitations, including preselection of "important" functionalities, a matrix intersection algorithm that is not easily generalized to substructures of three or more atoms or units, and no treatment of stereochemistry.

We propose a more general approach that seeks to identify a priori the common features of 3-D substructures of arbitrary size, independent of initial hypotheses about structural features responsible for activity, and that takes proper account of stereochemistry. Indeed, the approach is designed specifically to *generate* such hypotheses. Specifically, we propose a solution to the following problem. Given (1) a set of structures, all of which possess the "same" activity (the criteria for sameness of activity are the responsibility of the user), (2) a set of 3-D coordinates (x,y,z) for the location of every atom in every structure (these coordinates may have been obtained from X-ray crystallography or quantum mechanical or empirical force field energy minimization techniques), and (3) a set of user defined constraints (there may be none) as to what atom and substructural properties are important in the search for common substructures, find 3-D substructures of a given size or a largest size common to the set of structures, independent of whether or not atoms comprising the substructures are connected, while differentiating enantiomeric substructures if requested.

(2) METHOD

This problem is one of a large class of combinatorial problems for which all known algorithms are exponential in some key element, time or space, of the computations, in this case the number of possible substructures for a particular compound. Therefore, every effort must be made to constrain the search for common substructures and to employ efficient computational methods for making the search as efficient as possible. To illustrate important points in the process, we will follow a structure/activity problem through important phases of the search process. Some simple examples, which allow rapid and unequivocal testing of the algorithms, will also be presented.

The structure/activity problem involves warburganal (**1**)



and a number of related compounds [polygodial (**2**), muzigadial (**3**), and ugandensidial (**4**)] that possess antifeedant activity against African army worms. The compounds are also powerful helicocides (snail killers). In the interests of pest and disease control, it would be useful to know what substructural features are responsible for the activity because these features could then be incorporated into synthetic analogues on an industrial scale. This interest is due primarily to the helicocidal properties and the fact that some diseases, such as schistosomiasis, are spread by parasitic nematodes transmitted by snails. This structure/activity problem has been addressed in the literature,[10] so it will be possible to compare the results of the computer search with the results obtained by those researchers (who did not make use of computer assistance).

(2.1) Overview of Method. We summarize in Figure 1 the steps in our method for finding 3-D common substructures. The method is embodied in an interactive computer program. The input data consist of a file containing the structures to be analyzed, in the form of the x,y,z coordinates for every atom in every structure (at present, hydrogen atoms are ignored). A series of preprocessing steps convert the coordinate data into a distance matrix representation of the structures' atom types and interatomic distances. A hypothesized common 3-D pattern can be used as a starting point to begin the search process if desired.

The algorithm for identifying the common 3-D substructures has several strategies in common with the algorithm[3] for finding common topological substructures, although the method itself is quite different. In particular, one important strategy is to begin by finding all common substructures of the smallest size, beginning with the limiting, trivial case of one-atom substructures, if no larger starting point was specified. The algorithm proceeds by stepwise extension, or

Table I. Atom Types and Properties and Their Compacted Integer Descriptions

atom (of 1)	derived from				
	descriptor	aromaticity	degree	hybridization	type
C(8)	1041	non (=1)	n/a[a] (=0)	sp^2 (=16)	C (=1024)
C(1)	1049	non (=1)	n/a (=0)	sp^3 (=24)	C (=1024)
O(16)	2065	non (=1)	n/a (=0)	sp^2 (=16)	O (=2048)
O(18)	2073	non (=1)	n/a (=0)	sp^3 (=24)	O (=2048)

[a] Not applicable.

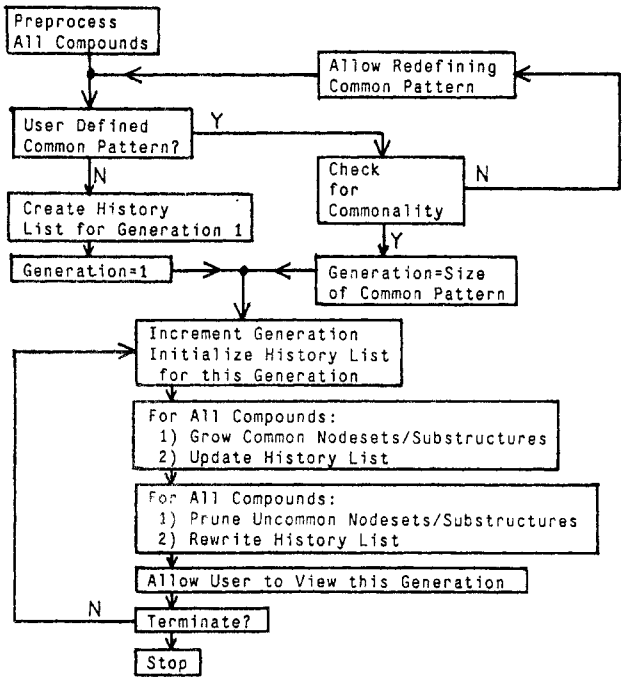


Figure 1. Flow chart for the search for common 3-D substructures.

"growing", by one atom to the next larger size, or "generation", checking for commonalities and continuing until terminating conditions are reached, for example, finding the largest common substructure or finding common substructures of a smaller, user-specified size.

A history list containing information about what substructures have been found in what structures is maintained throughout the procedure. This list is checked to ensure that substructures are common to at least a used-specified fraction of the structures. In addition, atoms found not to be part of any common substructure are flagged, because such atoms cannot be part of a common substructure of larger size.

Proper treatment of stereochemistry is a key feature of our method. Although the algorithm is made more complex by such treatment, it is a necessary aspect of this, or any other, method that attempts to correlate structure with activity because of the essential role of stereochemistry in many interactions of structures with the surrounding world. In the following sections we will discuss key elements of our method.

(2.2) Preprocessing. In order to search for common three-dimensional substructures, the x, y, and z coordinates of every atom in every compound must be available. Such information could come from X-ray crystallographic studies. If these data are not available, they could be obtained from a molecular modeling program. We have put together a system which will take a connection table (augmented with stereochemical descriptors) and, using tables of standard bond lengths and angles, generate an initial set of coordinates for all atoms in a structure.[11] These coordinate data are then refined by using them as the starting point for the MM2 energy minimization program,[12] resulting in a final set of coordinates corresponding to the nearest local minimum strain energy. Since, to our knowledge, X-ray studies have not been done on compounds **1–4**, it was necessary to use this approach to obtain coordinate data.

With the coordinate data in hand, it is necessary to form data structures which contain all data necessary for future processing steps in a convenient and easily accessible form. We have chosen to transform the coordinate space representation (x,y,z values) of each compound into a representation in distance space. This transformed representation contains all interatomic distances stored in an n by n array commonly called a distance matrix, where n is the number of atoms. For the 3-D search process, connectivity is unimportant and is disregarded (although connectivity tables are available). The distance matrix representation has three primary advantages: (1) it is an orientation-independent representation of a structure, (2) the necessary computations are simpler to carry out in distance space than in coordinate space, and (3) the distance matrix is one form of a special kind of graph called a "general graph"[13] in which every atom can be considered connected to every other atom. (In our case, the general graph is augmented with atom ("node") properties, and the "connections" ("edges") are augmented with distance information.) The general graph is attractive because (1) it has chemical significance in that any combination of atoms, whether or not actually connected in the original structure, can participate in a common substructure and (2) the characteristics of general graphs simplify subsequent computations, particularly construction of canonical names used in the comparison of substructures.

Two transformations are applied in the process of determining distance matrices for a set of structures. These transformations (1) reduce the storage space for atom names and (2) formally define the sets of distances that the program will consider to be the same.

(2.2.1) Description of Atom Names. We have chosen the following four descriptors as important in characterizing atoms: (1) atom type (i.e., C, N, O, etc.), (2) degree (i.e., quaternary, tertiary, secondary, or primary), (3) hybridization (i.e., sp^3, sp^2, or sp), and (4) aromaticity (i.e., aromatic or nonaromatic). Any, or all, of the four descriptors can be generalized to "not applicable", i.e., any atom type, or any degree, or any hybridization, or any aromaticity. The choice of descriptor will depend on the application and the judgment of the analyst as to what atom features might be important.

The description of atoms is compressed into an integer form which retains all information about atoms and is simple to manipulate in the computer. The following formula is used to calculate the integer atom descriptors:

atom descriptor =
2^0(2 aromatic, 1 nonaromatic, 0 not applicable) +
2^3(3 sp^3, 2 sp^2, 1 sp, 0 not applicable) +
2^6(4 4°, 3 3°, 2 2°, 1 1°, 0 not applicable) +
2^9(9 I, 8 Cl, 7 Br, 6 F, 5 S, 4 O, 3 N, 2 C, 1 H, 0 not applicable)

For the warburganal example, Table I illustrates examples of atom descriptors, under the assumptions used in subsequent analysis that aromaticity, hybridization, and atom types are important but that degree is not.

```
5.05 *3          2.68 *1
5.04 *           2.67 *3
5.03 *           2.66 *3
5.02 *           2.65 *3
5.01 *           2.64 *
5.00 *           2.63 *
4.99 *33         2.62 *
4.98 *           2.61 *21
4.97 *2          2.60 *1
4.96 *           2.59 *132
4.95 *           2.58 *43
4.94 *1          2.57 *22
4.93 *           2.56 *213143
4.92 *           2.55 *4
4.91 *3          2.54 *221231
4.90 *           2.53 *31344323
4.89 *           2.52 *34421221142
4.88 *42         2.51 *431144231
4.87 *           2.50 *231
4.86 *1          2.49 *3
4.85 *           2.48 *21
4.84 *13         2.47 *1
4.83 *           2.46 *32
4.82 *
4.81 *2          1.57 *32
4.80 *           1.56 *341221
4.79 *           1.55 *21331343
4.78 *           1.54 *341141231242423
4.77 *3          1.53 *31224243
4.76 *           1.52 *11
4.75 *
4.74 *
4.73 *
4.72 *3
```

Figure 2. Histogram of three distance groups between 1049- and 1049-type atoms. The numbers along the axis of the histogram are real distances rounded to the nearest 0.01 Å. The numbered elements making up the histogram refer to occurrences of the indicated distance in the four compounds (**1–4**) under study.

(2.2.2) Description of Interatomic Distances. The next preprocessing step involves the transformation of real distances into integer representations. This transformation is useful because it saves time in comparisons of distances. A much more important reason is, however, that distances will be compared for their equality, or "sameness", in the subsequent search for common 3-D substructures, and we need the transformation described below in order to make rational comparisons. Distances between the same types of *bonded* atoms, e.g., C_{sp^3}–C_{sp^3}, will vary slightly within a single structure and throughout a set of structures. Distances between a given pair of nonbonded atoms in geometrically very similar structures will show somewhat larger variations. One cannot simply compare such distances for precise equality. It is also not sufficient to apply a small distance uncertainty as a basis of comparison, because, inevitably, another distance for the same atom pair in another structure will fall slightly outside the uncertainty limits when it should be considered the "same".

One suggested solution for this problem is to use the root-mean-square (rms) difference in distances between two sets of (usually three or more) atoms.[4] Here one must still associate an uncertainty with the rms distance, and this calculation on a larger set of atoms will mask the effect of one large discrepancy in distance if other discrepancies are small.

We have examined several approaches to this problem of distance comparison, none of which is completely satisfactory. The approach we have adopted examines the natural clustering of distances among atom pairs of the same type to set boundaries on the "sameness" of distances by using the following set of steps.

(1) For all compounds under study (**1–4**, in our example), collect all distances between identical pairs of atoms and order them on the basis of magnitude. This means that, for instance, all distances between 1049- (nonaromatic, sp^3 carbons; see the section on the description of atom names) type atoms and 1049-type atoms will be grouped together. Likewise, all distances between 1041- and 1049-type atoms will be put into another group, distances between 1041- and 2065-type atoms in still another, and so on. Order the distances within each group on the basis of magnitude.

Table II. Portion[a] of the List of Mean Distances and Their Integer Pointers for the Many Clusters of Distances

integer pointer	mean dist, Å	integer pointer	mean dist, Å
.	.	49	2.84
.	.	50	2.73
.	.	51	2.41
14	1.54	.	.
.	.	.	.
.	.	.	.
.	.	73	4.76
27	2.95		
28	2.52		

[a] This portion was selected because examples in the text will refer to these distances.

(2) Break up each of the ordered distance groups into subgroups or clusters. A user-defined tolerance is utilized to define what constitutes a cluster, usually requiring an intercluster gap of 0.09 Å. An example is provided in Figure 2. The cluster from 1.52 to 1.57 Å represents the distribution of directly bonded 1049–1049-type atom pairs. The cluster from 2.46 to 2.68 Å contains 1049–1049 atom pairs separated by approximately two bonds. The third cluster, from 4.72 to 5.05 Å, represents pairs of 1049-type atoms separated by much larger distances. The clusters of more widely separated pairs of atoms are understandably less dense. One can obviously partition this distribution more finely, but it is better to err on the side of broader distributions when initially searching for common substructures. Later, the distribution can be partitioned more finely, but at the risk of overlooking real similarities allowed by the flexibility of molecules in real life (see Conclusions).

(3) Check each cluster to see if representatives from at least a user-specified number of the compounds occur. This is the case for the clusters in Figure 2 in which we have required all structures (**1–4**) to possess a given substructure. Most commonly, all compounds must be represented in each cluster. The option of having fewer than all compounds present in a "good" cluster allows the program to explore common substructures that are found in fewer than all the compounds.

(4) Whether or not the cluster is "good", find its mean distance value, and enter the mean at the end of a list of mean distances obtained from other clusters in this or other distance groups. See Table II for a portion of such a list. The purpose of this step is simply to have an integer value (the integer pointer location of the mean distance in the list of Table II) to represent a distance.

(5) If at least a user-specified number of compounds are represented in the cluster, simply enter the distance list pointer into the appropriate locations in the distance matrices of the compounds exhibiting this distance type. This fills up the upper half of the distance matrices for all compounds involved. If insufficient numbers of compounds were represented, flag the pointer (negate it), and then enter it in the appropriate location in the distance matrix.

This simple method for clustering distances is attractive because, in effect, the data are used to determine natural gaps in the clusters. It is probable, however, that a diverse and numerous set of structures would be difficult to treat in such a simple way, because the clusters at larger distances (greater than 3.0 Å) would begin to overlap considerably. An alternate method has been proposed[14] that avoids such problems by giving a set of four names, instead of one name, to an atom pair/distance combination, where each name consists of a representation of the atom types as in our method, together

Table III. Partial Distance Matrix for Warburganal (1)

atom no.	atom no.													
	1	2	3	4	5	6	7	8	9	10	...	16	17	18
1	1049	14	13	12	13	9	23	23	13	14	...	48	40	−99
2	1.53	1049	14	13	12	7	22	22	9	13	...	42	−34	−93
3	2.50	1.53	1049	14	13	9	22	22	7	12	...	38	33	−92
4	3.00	2.56	1.54	1049	14	13	23	23	9	13	...	38	−35	−93
5	2.52	2.94	2.53	1.55	1049	14	28	27	13	14	...	42	41	−99
6	3.88	4.36	3.92	2.58	1.53	1049	30	28	12	13	...	40	42	−95
7	4.34	5.23	5.06	3.92	2.52	1.50	1041	65	28	27	...	73	78	−108
8	3.89	5.10	5.23	4.43	2.88	2.52	1.34	1041	30	28	...	78	81	−110
9	2.53	3.92	4.34	3.96	2.52	2.98	2.49	1.52	1049	14	...	51	−49	−101
10	1.54	2.57	3.01	2.67	1.55	2.52	2.86	2.57	1.56	1049	...	48	43	−100
11	3.05	4.51	5.34	5.18	3.87	4.31	3.68	2.51	1.52	2.52	...	82	80	−110
12	5.02	6.34	6.54	5.78	4.23	3.74	2.32	1.35	2.49	3.82	...	−76	82	−109
13	2.50	3.15	3.84	3.37	2.56	2.97	3.30	3.15	2.52	1.54	...	−46	42	−95
14	3.90	3.20	2.51	1.53	2.61	3.07	4.51	5.20	4.87	3.46	...	−34	33	−89
15	4.31	3.87	2.48	1.55	2.54	3.13	4.46	5.17	4.96	3.96	...	33	33	−90
16	3.31	4.78	5.75	5.88	4.69	5.35	4.78	3.56	2.42	3.31	...	2065	−86	−118
17	5.28	6.73	7.11	6.56	5.01	4.78	3.45	2.27	2.82	4.33	...	3.29	2065	−117
18	2.79	4.16	4.38	4.24	2.94	3.65	3.23	2.32	1.40	2.40	...	2.70	3.12	2073

with a representation of a distance *range*. The set of names corresponds to the atom pair in four overlapping distance ranges. Although this method multiplies the number of atom pairs to be considered, it is an effective way to answer questions about distance similarities within a computer program.

The distance matrix that results from applying the aforementioned processing steps is given in Table III for warburganal (**1**). (Distance matrices of compounds **2–4** have similar format.) In this table, the lower half (below diagonal) of the matrix is filled with the original interatomic distances. The diagonal elements of the distance matrix (Table III) have been used to store the descriptors of the corresponding atoms. The above-diagonal elements contain the integer pointers to the respective mean distances of the distance list (see Table II). These pointers will be used to represent the distances in subsequent calculations. Thus, the C(2)–C(3) distance of 1.53 Å (lower half of distance matrix in Table III) is represented as 14 (upper half of the distance matrix), which represents the cluster of distances between two 1049-type atoms whose mean is 1.54 Å (Table II). The C(9)–O(17) distance (2.82 Å) is represented as −49, corresponding to the mean cluster distance of 2.84 Å (Table II), with the negative sign indicating that the distance is observed in fewer than four of the compounds.

(2.2.3) Common Patterns. As mentioned in the overview of the method, one can hypothesize the 3-D commonality of a given substructure in the set of compounds and use this substructure as the starting point in the search for larger substructures. The program checks that the substructure is, in fact, in common. We have tested this approach successfully, although we do not present examples of it in this paper. This approach can be very useful simply for testing hypotheses about 3-D structural similarity of a set of structures or for using known similarities (e.g., common ring systems) as an efficient starting point in searching for larger common features.

(2.3) Finding Common 3-D Substructures. The method for finding common 3-D substructures consists of three interrelated steps: (1) growing substructures, including formation of canonical, or unique, names for each substructure to allow rapid intercomparison of substructures; (2) updating and manipulating the history list of what substructures occur in what compounds; (3) removal of substructures found not to be in common and flagging of atoms that cannot participate in substructures of larger size as constraints on the next step of growing substructures.

Some definitions are important to the following discussion. The term *node* is used interchangeably with atom. The term *nodeset* is used throughout to refer to a set of specific, numbered atoms in a specific compound. For example, for warburganal (**1**), (11 16) is a nodeset consisting of nodes C(11) and O(16). The term *substructure* refers to a set of atom and distance descriptors for a set of atoms collected into a canonical, numbering-independent name by procedures discussed later. *Pruning* refers to elimination of one or more alternatives from further consideration. This term derives from the analogy of the tree-like form of the sets of nodesets (e.g., Figure 3).

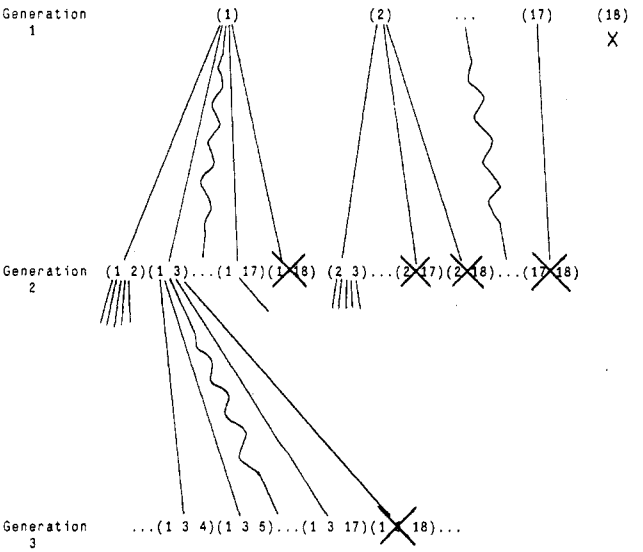


Figure 3. Growing of nodesets for warburganal (**1**). Note that generation 1 was created from *all* nodes/atoms in the compound. Generation 2 is grown by taking nodes from generation 1 and appending higher numbered nodes. Some nodesets of generation 2 are not considered since the distance between the nodes (see Table III) has been flagged as negative.

(2.4) Growing Substructures. The growing process requires that there be parent nodesets available. These nodesets can come from a previous generation (if one exists), from user-specified patterns (if the user has done so, this also affects the growing procedure, as will be discussed later), or, if neither of the preceding are available, from a routine that simply generates all nodesets of size one. Figure 3 illustrates part of the growing process for warburganal (**1**). In this figure, since there was no previous generation (hereafter referred to as the parent generation), the first generation was created from all nonhydrogen atoms in the compound. From this parent generation, an offspring generation was created by taking the nodesets from the parent generation and appending all nodes of higher numbering. This is done to avoid generating du-

plicate nodesets. If there were no restrictions on the node to be added, the nodeset (1 2) would be generated twice; once when node 2 was added to nodeset (1) and a second time when node 1 was added to nodeset (2). Thus, as shown in Figure 3, nodesets (1 2), (1 3), ..., (1 18) will be generated from nodeset (1), nodesets (2 3), (2 4), ..., (2 18) will be generated from nodeset (2), and so on. As each new nodeset is generated, the distances between the new node and the old nodes are checked to see if they are allowed (i.e., they have not been previously flagged as negative from the initial clustering of distances or subsequently from determination that the atom (and thus its distances to other atoms) cannot be part of a larger substructure). If the distance is not permitted, the nodeset is immediately discarded. For example, since all distances to O(18) have been flagged (Table III; sp^3 oxygen distances to all other atoms cannot be common because **2** possesses no sp^3 oxygens), all nodesets of size two or larger cannot contain node 18. Such nodesets are immediately discarded as shown in Figure 3.

In actual practice, the process of growing nodesets (and substructures) to the next largest size is carried out in parallel for all structures in the set. The results from a particular generation are used to constrain the growing procedure for the next generation in order to conserve time and space. This is accomplished by using the fact that a common substructure must appear in all (or a user-specified fraction) of the compounds.

The nodesets of Figure 3 and those obtained from structures **2–4** must be compared at every generation in order to determine what substructures are in common. In order to relate nodesets to substructures, we form a canonical, or unique, numbering-independent name for a nodeset and use this name, which includes a representation of all atom names and their interatomic distances (and stereochemistry, if desired), to represent a substructure. (The canonical naming procedure is described in a subsequent section.)

(2.5) History List. The commonality of substructures for the trivial, limiting case of one atom is considered implicitly in the process of constructing the distance matrices. Growing substructures of size two, and all subsequent sizes, is carried out for each structure in turn. For the first structure, nodesets are created by the growing process, and substructures are formed by the canonical naming procedure. At this point, we have no idea as to which will be in common, so all substructures and their corresponding nodesets are put in a history list.

For the second and succeeding compounds, each newly formed substructure is checked against the history list. If substructures are allowed that occur in less than all the compounds, checking is deferred until enough structures have been processed. The criterion is to check if at least [numcpd + 1 − incpd] compounds have already been processed (numcpd represents the number of compounds under study, and incpd is the number of compounds a substructure must occur in to be considered "common"). Thus, if numcpd = 4 and incpd = 4, as in our example, the substructures from the first compound will immediately be put in the history list, while those from the second through fourth compounds will be checked first. Pruning will take place during this step if those substructures that are checked are found to occur in less than [all − (numcpd − incpd)] of the compounds already processed. In our case, for structures **2–4**, if a substructure is formed which is not common to all preceding structures, it is immediately discarded. An example of this is shown as "intermediate" pruning in Figure 4.

In Figure 4, substructures A–C for structure **1** are added to the history list without checking. Structure **2** yields substructures A, B, and D. Substructure D and its nodeset can be discarded immediately because this substructure was not

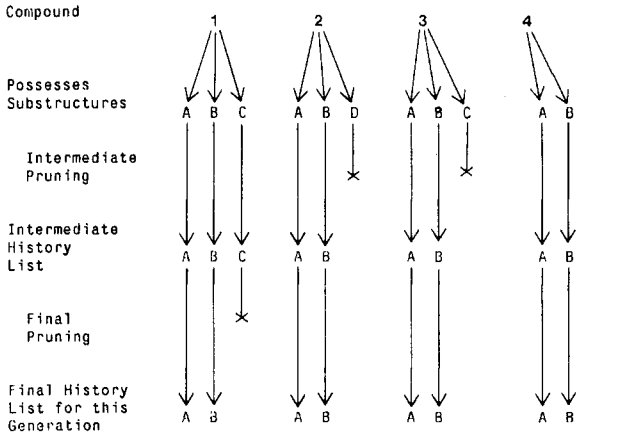


Figure 4. Example of intermediate and final pruning of a history list, assuming that substructures are required to be common to all structures.

present in structure **1**. (Note that the program could backtrack at this point and also delete substructure C from compound **1**, but we have chosen to defer this until a final pruning step.) For similar reasons, substructure C from **3** can be immediately discarded.

In the final pruning step, the history list is examined, and those substructures which are not present in enough (as defined by the user) of the compounds are discarded. This step removes substructure C from the history list as shown in the final pruning step of Figure 4. Subsequently, the list of nodesets is also examined, and those nodesets not included in surviving substructures are deleted. At the same time any interatomic distances not represented in any substructure are flagged negative in the distance matrix for the structure and, thus, are unavailable for future generations. A revised set of nodesets (and substructures) is obtained by this procedure, along with revised distance matrices which reflect the fact that some distances (and atoms) are no longer used. These manipulations reduce the number of nodesets and substructures to be considered during succeeding generations, thus saving time and storage space.

If a user-defined pattern has been used to start the growing process, some modifications have to be made to the above, although the basic steps are still the same. These modifications are as follows: (1) all the compounds in the study are examined to ensure that the user-defined common pattern is indeed common (three-dimensionally) to all compounds; (2) instead of beginning from generation 1, the starting point is that of the user-defined pattern; (3) when offspring nodesets are generated from a parent nodeset, only nodes of higher number than the highest node number in the parent nodeset *exclusive of the nodes in the defined common pattern* are considered. Any small increases in time and space are more than offset by the head start given by the user-defined pattern.

(2.6) Terminating Conditions. We have referred to our program as being interactive in nature because it is always up to the user to choose what to do next. This choice will generally be one of three: (1) view the common substructures for the current generation; (2) grow a next generation; (3) terminate the growing process (out of choice or because there are no common substructures at the current generation).

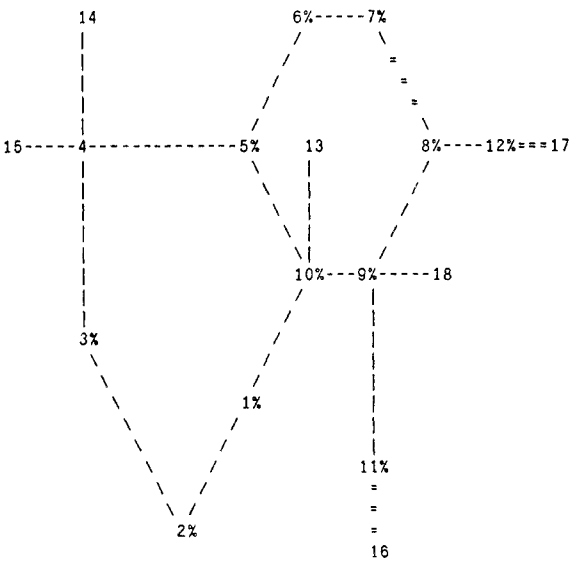
Viewing of the common substructures takes the form of a simple list of the substructure types along with the representative nodesets (as shown in Figure 5 for the common stereometric substructures of size 11 in the warburganal (**1**) example; see the Results section) or a drawing (atom named or numbered) of a specific nodeset (see Figure 6). The list is the shortest way of noting all the substructural types and how they occur in each structure. It is likely that some will be of more interest than others, and it is these latter sub-

```
>show nodesetsfor all
Substructure type 1
POLYGODIAL          1 2 3 5 6 7 8 9 10 11 12
MUZIGADIAL          1 2 3 5 6 7 8 9 10 11 12
WARBURGANAL         1 2 3 5 6 7 8 9 10 11 12
UGANDENSIDIAL       1 2 3 5 6 7 8 9 10 11 12
4 representative nodesets.

Substructure type 2
POLYGODIAL          1 2 3 5 6 7 8 9 10 11 16
MUZIGADIAL          1 2 3 5 6 7 8 9 10 11 16
WARBURGANAL         1 2 3 5 6 7 8 9 10 11 16
UGANDENSIDIAL       1 2 3 5 6 7 8 9 10 11 16
4 representative nodesets.
```

Figure 5. List of common stereometric substructures and nodesets for generation 11 of the warburganal (**1**) example.

```
>draw n (5 6)
NODESET 5 COMPOUND: WARBURGANAL
```




```
NODESET 6 COMPOUND: WARBURGANAL
```

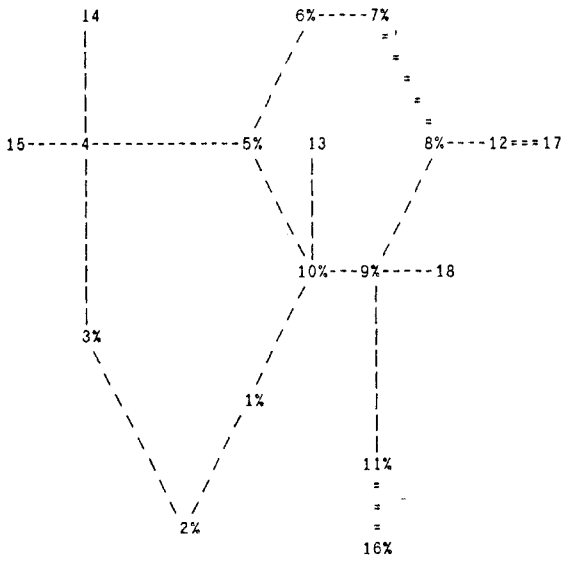


Figure 6. Atom-numbered drawings of the warburganal (**1**) nodesets representing the two substructural types indicated in Figure 5. The atoms making up the nodesets are tagged with a "%" marker.

structures and nodesets that can be drawn.

Growing of a next larger generation simply requires issuing the command to "grow". Having done so, the program proceeds to grow all common substructures of the next larger size (illustrated in Figure 7). If new common substructures are found, the numbers of substructures and representative nodesets are given; if none are found, that information is reported, as seen in the results of attempting to find larger common

```
>grow
Generating substructures of size 11
PROCESSING POLYGODIAL
PROCESSING MUZIGADIAL
PROCESSING WARBURGANAL
PROCESSING UGANDENSIDIAL
2 substructures grown.
8 representative sets of nodes.
>grow
Generating substructures of size 12
PROCESSING POLYGODIAL
PROCESSING MUZIGADIAL
PROCESSING WARBURGANAL
PROCESSING UGANDENSIDIAL
No substructures grown.
>exit
```

Figure 7. Illustration of the program's attempt to find common stereometric substructures of size 11 and 12.

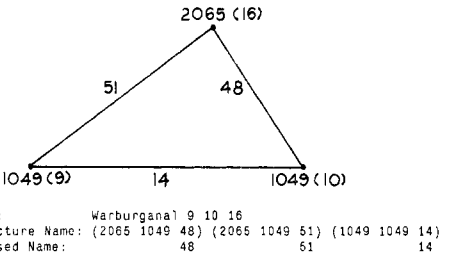


Figure 8. Steremetric canonical name for the nodeset consisting of atoms 9, 10, and 16 in warburganal (**1**).

substructures of size 12 in Figure 7.

Termination of the growing process is the user's choice. This will generally occur when nothing is found to be in common, when interesting substructures are found in common, or when the interactive session is interrupted. Information about a session is stored in files, so a new session can be begun from the point at which the old session was interrupted.

(3) CANONICALIZATION OF SUBSTRUCTURES

We deferred discussing the formation of canonical names for substructures to simplify the description of the algorithm. In this section we discuss this absolutely essential aspect of our method, because a procedure to name substructures in a unique way is required in order to compare them with one another.

We use two forms of canonical descriptors of 3-D substructures. When a user chooses not to differentiate enantiomeric substructures, we form *stereometric*[14] canonical names that retain all geometric information except that necessary for the differentiation. If enantiomers are chosen to be distinguishable, we form *stereochemical* canonical names.

(3.1) Stereometric Canonical Names. The design of the stereometric canonical name is based on the fact that any nodeset is an undirected general graph.[13] This means that atoms *n* and *m* are simply "connected" in the general graph, and the connection implies no direction from atom *n* to atom *m*. Thus, the atoms and their connections can be described by giving the descriptors of the two atoms involved, along with the integer representation of the distance between them. For this triple (atom *n* descriptor, atom *m* descriptor, distance (*n*,*m*), an ordering is imposed wherein the atom descriptors are ordered by decreasing magnitude, followed by the distance descriptor. Thus, referring to the distance matrix for warburganal (**1**, Table III), the canonical name for the two-atom nodeset consisting of atoms 9 and 16 is (2065 1049 51). In fact, because of the manner in which the real distances were transformed (see the section on Preprocessing), the integer distance descriptor implicitly contains the atom descriptors, thus resulting in a more compact version of the canonical name, simply 51.

For our stereometric canonical naming scheme, it is necessary to describe *all* interatomic connections for nodesets of size two or greater. Consider the three-atom nodeset 9, 10,

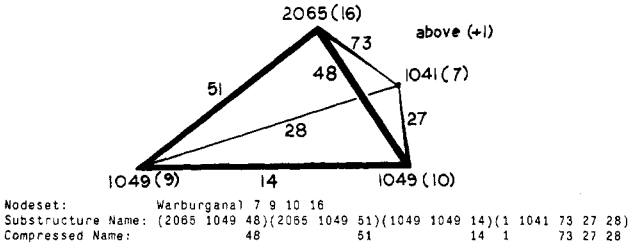


Figure 9. Stereochemical canonical name for the nodeset consisting of atoms 7, 9, 10, and 16 in warburganal (**1**).

16 for warburganal (**1**) shown in Figure 8. For these three atoms, there are three connections to be described. In general there will be ${}_nC_2$ such connections, where n is the number of atoms in the nodeset. This number will be of the order of n^2, so that large nodesets will require large canonical names.

In a first step, the ordered triples describing all connections are generated. Referring to the distance matrix of Figure 3 we see that the triples are (1049 1049 14) between atoms 9 and 10, (2065 1049 51) between atoms 9 and 16, and (2065 1049 48) between atoms 10 and 16. In a second step this collection of triples is ordered. In theory, any ordering scheme would suffice, as long as it was used consistently throughout. By our choice of ordering rules, though, names will tend to have their more unique elements toward the beginning of the name. The overall effect will be that in later stages, where it is necessary to compare canonical names (see the History List section), those names which are different will be noted as such after comparison of only a few elements, thus keeping the total time for such comparisons to a minimum. The order is based on the magnitudes of the elements of the triples. The rule is first to consider the magnitude (largest) of the first element, then the magnitude (largest) of the second element, and then the magnitude (smallest) of the third. Thus, the canonical name of the three-atom nodeset of Figure 8 is (2065 1049 48) (2065 1049 51) (1049 1049 14). Again, because the distance descriptors implicitly carry atom descriptors, the reduced canonical name (48 51 14) is obtained. A standard sorting routine ensures that the time required to order the list of triples is as small as possible, since for large nodesets the number of triples will be quite large.

For nodesets of size four (atoms) or greater, stereometric canonical names are created in the same way. If one chooses to differentiate enantiomeric (mirror-image) cases, the canonical naming procedure is modified both to take this into account and to reduce the amount of computer memory required to store the name.

(3.2) Stereochemical Canonical Names. Stereochemistry often plays an important role in determining the activity of a compound. However, in transforming the structural representation from x,y,z coordinates to distance space, we have discarded stereochemical information. (The substituents about the R and S forms of a given chiral center have the same distance relationships among themselves and the chiral center.) Consideration of stereochemistry must take place when sufficient atoms are included in a substructure to differentiate enantiomeric substructures. In general, this occurs with substructures containing four or more atoms, except when the substructure exhibits some form of symmetry. The ramifications of the symmetry will be discussed subsequently.

The procedure we use to derive the stereochemical canonical name of a substructure is as follows: (1) given a set of four or more atoms, choose three according to a number of rules to define a reference plane; (2) use the atomic coordinates and the right-hand rule to define what constitutes above and below the plane (see Figure 9); (3) next, for every other atom (not making up the planar three) in the structure, construct a five-element group containing the stereochemical descriptor (whether the atom is above or below the reference plane), the atom descriptor (discussed previously), and the distances to the three atoms used to define the reference plane; (4) finally, order the collection of groups, and append this list to the list formed from the stereometric canonical name derived from the three reference points.

As an example, consider the nodeset (shown in Figure 9) formed from atoms 7, 9, 10, and 16 of warburganal (**1**). In order to find a set of three reference points with which to define a reference plane, we must examine all the subsets of three atoms (in this case ${}_4C_3 = 4$; in general ${}_nC_3$ where n is the number of atoms in the nodeset) to see which satisfy the following criteria: (1) The set of three atoms must have the greatest (according to ordering rules as given in the Stereometric Canonical Names section) canonical name. (2) The canonical name must be unique (upon consideration of the stereometric canonical names of all three atoms nodesets within this nodeset). If the name is not unique, it is because some form of symmetry exists. This prevents choosing the three reference points, and thus defining the reference plane. (3) The stereometric canonical name must be made up of three unique parts. Otherwise it is not possible to define what is "above" and what is "below" the reference plane. (4) The three atoms must not be colinear so that it is possible to define a reference plane.

If the above criteria are not met, it is not possible to give the substructure a stereochemical canonical name. In such a case, the stereometric canonical name is generated and used to describe the substructure in question.

In considering the set of four atoms in Figure 9, a three-atom subset giving rise to the largest stereometric canonical name would necessarily involve atom 16 (descriptor 2065) and atoms 9 and 10 (both having descriptors of 1049). It is evident that the canonical name for this set of three atoms will be unique, that the three portions of the stereometric canonical name will be all unique (see the previous section for the name), and that the three atoms will not be colinear. Thus, the reference set consists of atoms 9, 10, and 16.

To define the sense of the reference plane (i.e., up and down), we use the stereometric canonical name of the three-atom reference set, in this example atoms 9, 10, and 16. The condensed name (48 51 14) is based on distances between atoms 10 and 16, 9 and 16, and 9 and 10, respectively. We assign priorities to these atoms based on the following rules. The atom common to the first two distance descriptors, atom 16, is assigned the highest priority, the atom common to the first and third descriptors, atom 10, is assigned next highest priority, and the remaining atom, atom 9, is assigned last priority. The right-hand rule is then used to define the above-plane direction. Figure 9 shows the atom 16, 10, 9 plane, along with what has been determined to be the above (+1) direction.

With an oriented reference plane defined, we can proceed to describe the locations of all other points with respect to the reference plane and points. In this case, only atom 7 remains to be described. It is above the plane (+1 descriptor), has an atom descriptor of 1041, and is at distances of 73, 27, and 28 from the ordered reference atoms (refer to the distance matrix of Figure 3). Thus, the pentuple for atom 7 is (1 1041 73 27 28). If there were other atoms, we would construct pentuples in a like manner and then order the pentuples according to a well-defined ordering scheme; vis, two or more pentuples are ordered based on the following set of priorities: stereochemical descriptor ($-1 > 0 > +1$), atom descriptor (highest magnitude has priority), and distance descriptors (lowest magnitude has priority). Pentuples are compared by pairwise comparison of their constituent elements. Priority of a pentuple depends on the priority of the elements at the first point of difference.

Table IV. Space Requirements in Computer Words for Storing Stereometric and Stereochemical Canonical Names for Nodesets of Increasing Sizes

size of nodeset	2	3	4	5	6	7	8	...	11	...	n
stereometric	1	3	6	10	15	21	28	...	55	...	${}_nC_2$
stereochemical			7	11	15	19	23	...	35	...	$3+4(n-3)$

Table V. Comparison of Numbers of Unique Substructures, Numbers of Nodesets, and the Corresponding Combinatorials for 5

generation	2	3	4	5	6	7	8
substructures	3	3	6	3	3	1	1
nodesets	28	56	70	56	28	8	1
${}_8C_{gen}$	28	56	70	56	28	8	1

For this example, all that remains is to append this pentuple to the stereometric name of the reference set of atoms to get the stereochemical canonical name of this substructure: (2065 1049 48) (2065 1049 51) (1049 1049 14) (1 1041 73 27 28). The condensed name is then (48 51 14 1 73 27 28) as shown in Figure 9.

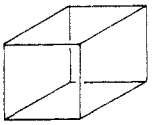
For a general nodeset of size n, the amount of space required to store the name will be $3 + 4(n - 3)$ computer words, which for n greater than 6 will always be less than the size required to store stereometric canonical names. An important point is that the space requirement for the stereochemical canonical name is linear, not exponential, in the number of atoms, n, unlike the stereometric canonical name. See Table IV for a comparison of space requirements for storing stereometric and stereochemical canonical names.

A drawback to the use of stereochemical canonical names is the dependency of the names on coordinate data for determining what is above and below the reference plane. Considering that the distances used actually represent distance ranges, some of which are relatively wide, assignment of correct stereochemical parity for atoms close to the reference plane is problematical. Two geometrically very similar substructures may yield different parities and will be viewed by the program as different.

(4) RESULTS

In previous sections, various aspects of the 3-D search have been illustrated. Most of the examples were obtained by using the realistic example of **1**–**4**. To verify that the search process discovers the correct numbers of substructures and nodesets, we have examined several much simpler examples. We first present some results from these tests and then return to common substructures for **1**–**4**.

(4.1) Stereometric Substructures. One of the key features of the program is its ability to grow and describe substructures in three dimensions. Both the growing process and the method for canonically naming substructures have been described earlier, so we now present a simple example of several tests used to confirm that the program works on easily verifiable examples. In this case, consider a simple cube. Its many symmetries force the program to deal exclusively with stereometric canonical names. Indeed, if this example is run allowing differentiation of enantiomers, only stereometric names are generated. Table V contains the resulting numbers of substructures and nodesets obtained by starting from size 2 and growing to the maximum size 8 for a cube (**5**).



5

We see that the total number of nodesets grown is simply the number of combinations of the number of atoms, 8, taken

Table VI. Comparison of Numbers of Unique Stereometric Substructures of Varying Sizes with the Corresponding Combinatorials for Structures **6** and **7**

generation	2	3	4	5
substructures	10	10	5	1
${}_5C_{gen}$	10	10	5	1

Table VII. Comparison of Numbers of Common Stereochemical Substructures of Varying Sizes with the Corresponding Combinatorials for Structures **6** and **7**

generation	2	3	4	5
substructures	10	10	0	0
${}_5C_{gen}$	10	10	5	1

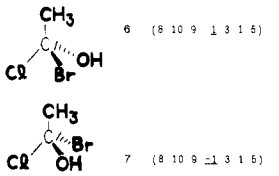
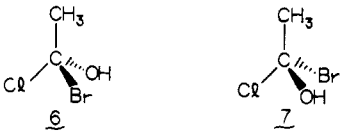


Figure 10. Stereochemical canonical name for the highlighted atoms in **6** and **7**.

n at a time, where n is the generation. Verifying that the number of substructures and nodesets are correct takes a bit of thought, but the numbers are, indeed, correct.

A second example is presented in which there are no symmetries. In this case, the two structures, **6** and **7**, differ *only*



in that they are mirror images of each other. From a stereometric point of view, they should be found to be identical out to the maximum size, and the number of substructures should be a simple combinatorial.

Indeed, as seen in Table VI, the number of unique substructures for each generation is as predicted.

(4.2) Stereochemical Substructures. The results obtained for substructures of **6** and **7** considering stereochemical substructures are summarized in Table VII.

In the first generation in which stereochemistry is considered (generation 4, Table VII), there are found to be *no* common substructures, the intuitively obvious result. The program's detection of nonequivalence is through comparison of the stereochemical canonical names, which are shown for one substructure in Figure 10.

The substructures are mirror images, and the derived stereochemical canonical names differ only in the parity of the stereochemical tags, which are underlined in Figure 10.

(4.3) Application to Structure/Activity Relationships. Simple examples have served to demonstrate various aspects of the program, to show that correct numbers of substructures are grown, and to show that the correct substructures are grown. We now continue the structure/activity problem which was introduced earlier. Recall that all four compounds **1**–**4** have been preprocessed with the following conditions: (1) atom type, hybridization, and aromaticity are used as atom descriptors; (2) a gap of 0.09 Å was used for clustering of distances during preprocessing. The following sections show the results of our search for common 3-D substructures considering stereometric and stereochemical substructures.

(4.3.1) Growing Stereometric Substructures. Without differentiation of enantiomers, the program determines the numbers of nodesets and 3-D stereometric substructures given in Table VIII.

Table VIII. Numbers of Nodesets and Common Stereometric Substructures Obtained for Each Successive Generation for Structures 1-4[a]

generation	common stereometric substructures	nodesets
2	31	499
3	161	1567
4	440	3084
5	712	3975
6	756	3669
7	561	2451
8	289	1212
9	101	407
10	21	84
11	2	8
12	0	0

[a] A total of 26 min of cpu time was used.

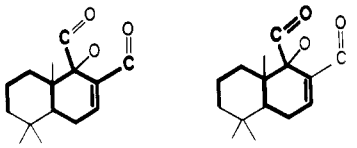


Figure 11. Two common stereometric substructures of size 11, illustrated by using warburganal (**1**). Bonds connecting common substructural atoms have been highlighted for clarity but are *not* formally considered part of the substructure because connectivity is not a consideration in the search for 3-D common substructures.

These data are interpreted as follows. For each generation, the generation number refers to the size of the substructure. Thus, for generation 2, or two atom substructures, there are 31 different, common, 3-D stereometric substructures, all of which are present at least once in each of **1–4**. The 499 nodesets indicate that the 31 common substructures occur a total of 499 times in **1–4**. A detailed breakdown of the multiple occurrences of each substructure cannot, of course, be obtained from the table, but these data are available as output from the program. In successive generations, the number of common substructures increases rapidly but begins to decrease at generation 7, as more and more distances in the distance matrices are flagged negative (not in common) and are not carried on to the next generation.

The largest common stereometric substructures possess 11 atoms, and there are two such substructures, occurring once in each structure (Table VIII). These substructures are illustrated in Figure 11 by using warburganal (**1**) as an example. The second substructure in Figure 11 must be regarded as an artifact of the rigid models obtained from the molecular modeling program, because, although rotation about the C(9)–C(11) bond is restricted, it is unlikely that O(16) would be so conformationally locked in space.

In general, one can hypothesize that the larger common 3-D substructures are responsible for the characteristic activity of a set of structures. Such hypotheses should, however, be tempered by knowledge of the variety of structural forms used in determining common substructures. In this case, we have probably overdetermined the pharmacophoric pattern postulated to be required for activity (see below), because structures **1–4** possess a common ring system.

In spite of the similarities in the substructures shown in Figure 11, there is no common substructure of size 12, for example, one which includes both O(16) and C(12) of **1**. Examination of the distance matrices for all four compounds under study revealed that the interatomic distance between the carbonyl carbon of the enal and the carbonyl oxygen of the other aldehyde group was not similar enough in all four compounds (3.80 Å in **1** and **2**, but 3.69 in **3** and 3.70 Å in **4**). This discrepancy in distances, under the conditions used in our analysis, is sufficient to rule out the participation of this pair of atoms in common substructures.

Table IX. Numbers of Nodesets and Common Stereochemical Substructures Obtained for Each Successive Generation for Structures 1-4[a]

generation	common stereochemical substructures	nodesets
2	31	499
3	161	1567
4	432	2458
5	544	2558
6	464	1989
7	276	1143
8	117	472
9	33	132
10	5	20
11	0	0

[a] Enantiomeric substructures were differentiated in this example. A total of 38 min of cpu time was used.

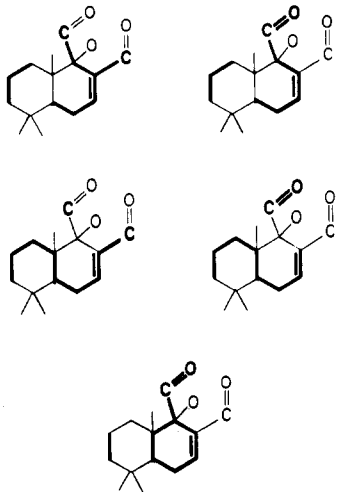


Figure 12. Five common stereochemical substructures of size 10, illustrated by using warburganal (**1**). Bonds connecting common substructural atoms have been highlighted for clarity but are *not* formally considered part of the substructure because connectivity is not a consideration in the search for three-dimensional common substructures.

(4.3.2) Stereochemical Substructures. We summarize in Table IX the numbers of common stereochemical substructures and nodesets found by our program when enantiomers are differentiated.

Given the stereochemical similarity of common features of **1–4**, we expect to obtain very similar results comparing Tables VIII and IX, as is the case. Most of the differences between the tables are due to the sensitivity of formation of stereochemical canonical names to the preprocessing steps and the choice of the reference plane alluded to earlier. As an example, the maximum common substructures are only of size 10. The five detected are illustrated in Figure 12, using the structure of warburganal (**1**). These substructures capture the essence of the common substructures shown in Figure 11. Failure to proceed to size 11 is due only to the sensitivity of the stereochemical naming procedure and *not* to a differentiation between enantiomeric substructures. This example overemphasizes the difficulties inherent in characterizing stereoisomers. If enantiomers of **1–4** had been included in the analysis, the stereometric naming procedure would be incapable of differentiating many substructures, while the stereochemical naming procedure would easily perform the differentiation.

(4.3.3) Stereometric Substructures from a User-Defined Common Pattern. The previous two searches for common 3-D substructures resulted in the growth of large numbers (see Tables VIII and IX) of substructures and nodesets, particularly

Table X. Nodesets and Common Stereometric Substructures for 1-4 for Successive Generations when Starting from a User-Defined Pattern[a]

generation	common stereometric substructures	nodesets
4	1	4
5	6	35
6	23	120
7	45	218
8	52	235
9	36	147
10	13	52
11	2	8
12	0	0

[a] A total of 8 min of cpu time was used.

in the early generations. This was due both to the sizes of the compounds under study (**1–4**) and to their obvious similarities. This necessarily means that a large quantity of computer space will be required to store all relevant information and that a large amount of time will be required to carry out the growth through all generations. There is always an upper limit to the amount of available space, and an increase in time is always undesirable. One way to circumvent both of these problems is to have the user define a starting point. This starting point will be a substructure that he or she sees present in all compounds under study. This substructure will probably also represent a hypothesis about portions of the compounds responsible for activity. This bias will prevent the program from considering an exhaustive list of common substructures, carrying with it the possibility that one of those that is not considered is, in fact, responsible for the activity of the compounds under study. Even so, this approach represents an excellent way of quickly checking and extending hypotheses.

An examination of the structures of warburganal (**1**), polygodial (**2**), muzigadial (**3**), and ugandensidial (**4**) reveals that large portions of the compounds are in common. As an example, in order to give the search process a head start, yet not constrain it too severely, we selected the nodeset consisting of atoms 6–9 (the olefinic portion of the "B" ring) as potentially common *in a three-dimensional sense*. The program verifies that this nodeset is indeed common to all four compounds, begins at generation 4, and grows those nodesets and substructures of size 5 that contain that set of four atoms. This drastically reduces the numbers of substructures and nodesets generated (compare Table X with Table VIII), with a concomitant decrease in storage space used and run time.

The two common substructures of size 11 eventually found are the same as those found when the problem was started at generation 2 and allowed to continue until the substructures of maximum size had been generated (Figure 11).

In the literature[10] it was determined that the enal 9β-aldehyde moiety plays an essential role in the antifeedant activity. This was done by carrying out chemical conversions on polygodial (**2**) and testing the products for activity. Epimerization to give the 9α-aldehyde resulted in loss of activity, implying that a 9β-aldehyde was necessary. Reduction products (enediol and diol) and oxidation products (diacid and lactone) were found to be inactive, leading to the conclusion that the aldehyde groups and olefin were required. These moieties are present in one of the two common substructures found in the computer search (see Figure 11) and, of course, are present in substructures of smaller size together with other substructures that refer to common elements of the ring system itself. The requirement for these elements could only be determined by expanding the set of compounds to include other compounds with different ring systems. Thus, the search process that we have described has confirmed that the proposed active substructure[10] is, indeed, common in a 3-D sense to all four active compounds and thus could be responsible for the activities. Obviously, the results of the program could be used prospectively to propose substructures required for activity, which then could be confirmed by chemical conversions and biological tests.

(4.4) Conclusion. In designing a program for finding three-dimensional structural similarities, we have tried to include as much flexibility as possible and at the same time minimize time and storage requirements. Despite these efforts, an unconstrained problem, beginning from scratch and continuing to the maximum common substructure(s), is apt to consume a large quantity of time, and thus be essentially noninteractive. The solution is that only "small" problems (small numbers of compounds or small-sized compounds) should be run this way. Larger problems require some form of user constraint as to what appears to be common to all structures or a hypothesis of what the important portions of the structures are. Only in this manner can the exponential explosion of substructures at early generations be circumvented.

There are still some problems that remain to be solved satisfactorily. These all stem from the necessity of dealing with "raw" data (x, y, and z coordinates determined from X-ray studies or from modeling programs). For instance, the method of converting real distances into integer "pointers" requires more robust and general methods, as we discussed in the Preprocessing section. Another weak link in the program is the stereochemical canonical naming method for reasons discussed previously. It has been shown to work well on examples where the data (x,y,z coordinates) are of high quality and where the real to integer distance conversion step results in integer descriptors that define very narrow ranges of distances.

In spite of these acknowledged weak points, this program does represent one of the first solutions to the important problem of detection of common 3-D substructures. Although the combinatorial explosion for large numbers of substructures is not eliminated by this algorithm, it is at least reduced substantially by transformation of the problem from coordinate space to distance matrix representations, together with some clever computational tricks involved with using the history list to constrain the procedure.

Our approach has two methodological problems. The first is its implicit assumption that all compounds share the same activity. Thus, negative results, i.e., information about inactive structures, cannot be used to guide the search procedures except by manually giving the program hypotheses about common substructures. In real life,[10] such negative results are often critical to focussing on the essential structural elements required for activity. Related computational approaches are being developed to remove this methodological limitation.[14]

A second methodological problem is the program's view of structures as rigid entities. Although some conformational flexibility is considered by using clusters, or ranges, of distances, it would be a misuse of the program to apply it to comparison of conformationally flexible structures. Although some programs explore conformational space in structure comparison,[8] they require an initial hypothesis about a pharmacophore. A static representation of inherently dynamic structures that is useful for detection of common 3-D substructures is now being developed.[14]

EXPERIMENTAL SECTION

These programs are implemented in the ALGOL-like BCPL programming language[15] on a Digital Equipment Corp. KI-10 computer at the SUMEX-AIM facility at Stanford University. The programs are available through Molecular Design Ltd., Hayward, CA.

ACKNOWLEDGMENT

We thank the National Institutes of Health (Grant No. RR-00612-12) for their generous financial support. Computer resources were provided by the SUMEX facility at Stanford University under National Institutes of Health Grant RR-0785.

REFERENCES AND NOTES

(1) Part 44 of the series "Applications of Artificial Intelligence for Chemical Inference". For part 43 see: Lindley, M. R.; Shoolery, J. N.; Smith, D. H.; Djerassi, C., *Org. Magn. Reson.*, in press.
(2) Cone, M. M.; Venkataraghavan, R.; McLafferty, F. W. "Molecular Structure Comparison Program for the Identification of Maximal Common Substructures". *J. Am. Chem. Soc.* **1977**, *99*, 7668–7671.
(3) Varkony, T. H.; Shiloach, Y.; Smith, D. H. "Computer-Assisted Examination of Chemical Compounds for Structural Similarities". *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 104–111.
(4) Gund, P.; Wipke, W. T.; Langridge, R. "Computer Searching of a Molecular Structure File for Pharmacophoric Patterns". In "Proceedings International Conference on Computers in Chemical Research and Education"; Elsevier: Amsterdam, 1973; pp 5–33.
(5) Lesk, A. M. "Detection of Three-Dimensional Patterns in Chemical Structures". *Commun. ACM* **1979**, *22*, 219–224.
(6) Rohrer, D. C.; Perry, H. "FITMOL". In "Public Procedures: A Program Exchange for PROPHET Users"; Wood, J. J., Ed.; Bolt, Beranek, and Newman, Inc.: Cambridge, MA, 1978.
(7) Cohen, N. C. "Beyond the 2-D Chemical Structure"; In "Computer Assisted Drug Design"; Olson, E. C., Christoffersen, R. E., Eds.; American Chemical Society: Washington, DC, 1979; pp 377–381.
(8) Marshall, G. R.; Barry, C. D.; Bosshard, H. E.; Dammkoehler, R. A.; Dunn, D. A. "The Conformational Parameter in Drug Design: The Active Analog Approach"; In "Computer-Assisted Drug Design"; Olson, E. C., Christoffersen, R. E., Eds.; American Chemical Society: Washington, DC, 1979; Chapter 9, pp 205–226.
(9) Avidon, V. V.; Pomerantsev, I. A.; Golender, V. E.; Rozenblit, A. B. "Structure-Activity Relationship Oriented Languages for Chemical Structure Representation". *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 207–214.
(10) Nakanishi, K.; Kubo, I. "Studies on Warburganal, Muzigadial, and Related Compounds". *Isr. J. Chem.* **1977**, *16*, 28–31.
(11) Wenger, J. C.; Smith, D. H. "Deriving Three-Dimensional Representations of Molecular Structure from Connection Tables Augmented with Configuration Designations Using Distance Geometry". *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 29–34.
(12) Allinger, N. L. "Conformational Analysis. 130. MM2. A Hydrocarbon Force Field Utilizing V_1 and V_2 Torsional Terms". *J. Am. Chem. Soc.* **1977**, *99*, 8127–8134.
(13) Harary, F. "Graph Theory"; Addison Wesley: Reading, MA, 1971.
(14) Smith, D. H.; Carhart, R. E.; Crandell, C. W.; Venkataraghavan, R. "Constructive Perception of Shared Three-Dimensional Substructures". "Abstracts of Papers"; 186th National Meeting of the American Chemical Society, Washington, DC, Aug 28–Sept 2, 1983; American Chemical Society: Washington, DC, 1983; CINF No. 3.
(15) Richards, M.; Whitby-Stevens, C. "BCPL – the Language and its Compiler"; Cambridge University Press: Cambridge, 1979.

Carbon-13 Nuclear Magnetic Resonance Spectral Interpretation by a Computerized Substituent Chemical Shift Method†

H. N. CHENG* and S. J. ELLINGSEN

Hercules Incorporated, Research Center, Wilmington, Delaware 19899

Received December 30, 1982

A FORTRAN computer program (called CSHIFT) is developed for the rapid estimation of the ^{13}C NMR chemical shifts of aliphatic organic compounds. The method is based on additive ^{13}C shift relationships, using empirical substituent chemical shift parameters. Examples are given that illustrate its use.

INTRODUCTION

It is generally recognized that ^{13}C NMR spectroscopy is a very powerful tool for organic structure determination. A major task in the interpretation of ^{13}C NMR spectra is to estimate the chemical shifts of compounds known or suspected to be present. Two approaches are generally used: (1) look up the chemical shifts in spectral libraries of either the compound in question or, if not available, compounds with similar structures; (2) calculate the ^{13}C shifts by using empirical substituent chemical shift rules.

For the first approach the spectral collections of Sadtler,[1] Bremser,[2] Breitmaier,[3] and Stothers,[4] among others, are very useful. In the last few years, many computer-assisted structure-determination methods have been developed.[5] Some of the earliest, the CNMR program[6] of Chemical Information Systems and its variants, have been generally available for several years. Recently the Stanford group has developed an array of sophisticated methods.[7] Several other groups are also very active in advancing this important area.[8-15]

In the second approach, there exist empirical rules such as those formulated by Grant and Paul,[15] Lindeman and Adams,[16] and Carman, et al.[17] for hydrocarbons, by Eggert and Djerassi[18] and Sarneski et al.[19] for amines, by Roberts[20] and Ejchart[21] for alcohols, and by Hagen and Roberts for carboxylic acids[22], along with numerous others observed for other functional groups.[23,24] Clerc and Pretsch have devised general additive rules for 28 functional groups.[25] Dubois has used a topological parameter to model the alkyl environment.[26] Levy and Nelson[27] and Ejchart[28,29] have proposed substitution methods whereby the ^{13}C shifts are first estimated for the hydrocarbons, and heteroatoms are substituted later. Although these rules have varying accuracy, they serve as good starting points for spectral interpretation, especially when simple analogues cannot be located in the spectral libraries. A drawback to this approach is that it is labor-intensive and occasionally prone to arithmetic error.

One way to facilitate the application of substituent chemical shift rules is to computerize them. One such effort was made by Clerc and Sommerauer.[30] In this work we have modified and extended the Clerc–Pretsch rules and computerized them using a different approach. Our program (called CSHIFT) was written in a high-level language (FORTRAN IV) and has many special features. It is applicable to aliphatic carbons carrying 30 functional groups including the 28 listed by Clerc and Pretsch.[25] It can also take care of alicyclic compounds, although its accuracy tends to be lower.

METHOD

In the Grant–Paul scheme,[15] the ^{13}C shifts are thought of as arising from empirical additive parameters that are char-

† Hercules Research Center Contribution No. 1762.

0095-2338/83/1623-0197$01.50/0