ACKNOWLEDGMENT

The support of this research through National Science Foundation Grant CHE-76-21295 is gratefully acknowledged. We also acknowledge partial support from U.S. Department of Energy, Division of Basic Energy Research, administered through R. S. Hansen, a grant to Ames Laboratory, Iowa State University.

REFERENCES AND NOTES

(1) On leave from Ames Laboratory, Iowa State University, Ames, Iowa 50011.

(2) For a recent survey, see R. C. Read and D. G. Corneil, *J. Graph Theory*, **1**, 339 (1977).

(3) L. C. Ray and R. A. Kirsch, *Science*, **126**, 814 (1957).

(4) J. E. Ash, "Connection Tables and Their Role in a System", in "Chemical Information Systems", J. E. Ash and E. Hyde, Ed., Wiley, New York, 1975, p 167.

(5) J. Figueras, *J. Chem. Doc.*, **12**, 237 (1972).

(6) G. Saucier, *R. I. R. O.* 5e annee, No. R-3 (1971) 31, MR 46.1654 (as quoted in ref 2).

(7) E. H. Sussenguth, *J. Chem. Doc.*, **5**, 36 (1965).

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

(9) M. Randić, *J. Chem. Inf. Comput. Sci.*, **18**, 101 (1978).

(10) L. Euler, *Comm. Acad. Sci. Imp. Petropol.*, **8**, 128 (1736); reprinted in N. L. Biggs, E. K. Lloyd, and R. Wilson, "Graph Theory 1736-1936", Clarendon Press, Oxford, 1973, p 3.

(11) A. V. Aho, *Acta Crystallogr.*, **33**, 5 (1977). E. L. Lawler, *Math. Centre Tracts*, **81**, 3 (1976). R. E. Tarjan, "Complexity of Combinatorial Algorithms", Computer Science Dept., Stanford University, Report STAN-CS-77-606, April, 1977.

(12) M. Randić, *J. Chem. Inf. Comput. Sci.*, **17**, 171 (1977).

(13) M. Randić, *J. Chem. Soc., Faraday Trans. 2.*, submitted for publication.

(14) M. Randić, *Chem. Phys. Lett.*, **58**, 180 (1978).

(15) M. Randić, G. M. Brissey, R. B. Spencer, and C. L. Wilkins, *Comput. Chem.*, in press.

(16) M. Randić, *Chem. Phys. Lett.*, **42**, 283 (1976); *Croat. Chem. Acta*, **49**, 643 (1977); *Int. J. Quant. Chem.*, in press.

(17) M. Randić, Proceedings of Bremen Conference: "Mathematical Structures in Chemistry" (June 1978, Bremen, Germany), to appear in MATCH (Informal Communications in Mathematical Chemistry, edited by A. T. Balaban, A. S. Dreiding, A. Kerber, and O. E. Polansky).

(18) Label A has to be counted with weight 2, because it represents two atoms.

Graph Theoretical Approach to Recognition of Structural Similarity in Molecules

MILAN RANDIĆ*[1a] and CHARLES L. WILKINS*[1b]

Department of Chemistry, University of Nebraska–Lincoln, Lincoln, Nebraska 68588, and Ames Laboratory, Iowa State University, Ames, Iowa 50010

Received June 8, 1978

In many applications, one is faced with the problem of identifying similar structural forms. This is usually performed in an intuitive manner and the outcome may be ambiguous. Here a well-defined approach for determining the degree of similarity among structures (molecular skeletons and more general graphs) is suggested. We select paths in a structure as the invariants upon which comparisons among structures should be based. For each structure, one first enumerates paths of different length and constructs a sequence of path numbers for the atoms (vertices). From such a list of atom codes, one can derive, by summing the contributions of individual atoms, a sequence of path numbers for a molecule (or a graph). The comparison of structures, thus, can be transformed into a comparison of *sequences* which are suitable for rigorous mathematical analysis. It is *assumed* that similar sequences imply similar structures, and, for selected examples, the validity of this assumption has been demonstrated. Molecules, generally considered similar, have been found to have similar sequences of path numbers, and molecules differing considerably in their connectivity show large differences in their path numbers. As an illustration of the concept of similarity based on path enumerations, we consider the problem of selecting from a set of structures those most similar to naturally occurring monocyclic monoterpenes. We used available computer-generated hypothetical monocyclic monoterpenes. The motive for this comparison is the assumption that potentially interesting skeletal forms should show considerable similarity to those found in natural structures. Such a technique may be of particular use in problems in chemical taxonomy.

INTRODUCTION

In many applications, one is confronted with the task of identifying a few structurally related systems among dozens or hundreds of others which are similar. One must begin by defining the concept of *similarity* which may apply to a selected property, to a dominant property, or to the overall features of a system (or objects) under examination. We will be concerned here with the *structural* parameters of molecules, each structure being specified by molecular connectivity (i.e., the adjacency or structural matrix).[2] In order to use this definition of similarity on a quantitative level, we require a measure of similarity and specified tolerance levels which can be used to categorize structures as similar or dissimilar.[3] For example, we may wish to review physical and chemical properties of alkanes (boiling points, solubility, vapor pressure dependence on temperature, viscosity, optical density, chromatographic retention volumes, density, etc.), in particular isomeric variations. To call two isomers similar (with respect to a single selected property, e.g., boiling point), we have to specify tolerance levels and then see if differences among the isomers fall within the specified interval. Objections to such an empirical approach mainly arise from its a posteriori character, which dictates that whatever conclusion is drawn is of limited guidance in a completely new situation. For example, similarity in boiling points and/or a few other thermodynamic properties cannot assist in predicting which members of a group of therapeutically useful compounds will show the optimum antihistaminic, anticholinergic, antipsychotic, antidepressant, analgesic, or anticonvulsive potency. Here, it would be more useful to have several representative compounds of each class (standards) and then to compare the candidate structure with the standards, using as many properties as thought to be pertinent. Since many molecular properties, and especially chemical or therapeutic activity, bear some relationship to chemical structure, studies of the similarity of *structures*, rather then *properties*, should be the first

© 1979 American Chemical Society

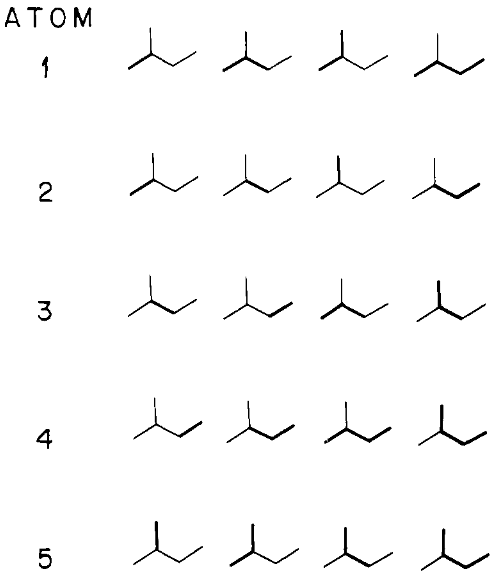


Figure 1. Paths for each atom of 2-methylbutane are depicted by thick lines. Notice each path occurs twice, once for each end atom.

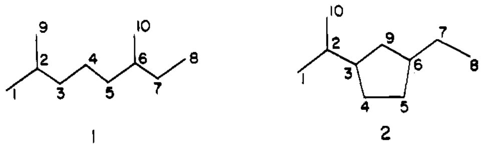


Figure 2. Carbon skeletons of an acyclic and a cyclic monoterpene.

priority. One of us has previously suggested an approach to such a comparison, using molecular transforms of candidate structures and applying pattern recognition methodology.[4] In the present paper, we consider an alternate approach to detecting structural similarity based on comparison of selected graph invariants representing partial skeletons. In particular, we find enumeration of all self-avoiding paths is a useful basis for characterization of similarity. A self-avoiding path is defined in a graph as an alternating sequence of vertices and edges which are adjacent (bonded) with no vertex appearing more than once. In Figure 1, all paths for each atom of the carbon skeleton of 2-methylbutane are diagrammed. Notice, each path occurs twice, once for each of the two terminal atoms. It will be demonstrated with selected representative examples that indeed molecules having approximately the same number of paths of different length (i.e., having a similar distribution of paths) also show the apparent structural similarity that an intuitive approach would suggest. Considering this, we hypothetize that path enumerations and path numbers preserve important structural elements and could provide a basis for *rigorous* comparison of structures. A measure of similarity follows from the count of differences in the number of paths of different length in the structures considered.

PATH ENUMERATION IN MOLECULAR GRAPHS

In acyclic graphs, any two vertices are connected by a single path. Enumeration of paths is straightforward: one simply considers each atom separately and forms a sequence of numbers, the first entry giving the number of paths of length one, the second entry giving the number of paths of length two, and so on. This is illustrated for a ten-carbon atom "head-to-tail" isoprenoid skeleton (Figure 2). The resulting sequences are given in Table I and will also be referred to as atom codes. An example of the atom path enumerations for the cyclic structure in Figure 2 appears in Table II. For both examples, only equivalent vertices (atoms) have identical path sequences. However, occasionally nonequivalent atoms can

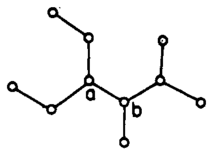
Table I. Sequences of Paths of Length One, Two, Three, etc., for an Acyclic System Containing a Head-to-Tail Isoprenoid Structure

atom	paths
1	1, 2, 1, 1, 1, 2, 1
2	3, 1, 1, 1, 2, 1
3	2, 3, 1, 2, 1
4	2, 2, 4, 1
5	2, 3, 2, 2
6	3, 2, 1, 1, 2
7	2, 2, 1, 1, 1, 2
8	1, 1, 2, 1, 1, 1, 2
9	1, 2, 1, 1, 1, 2, 1
10	1, 2, 2, 1, 1, 2

Table II. Sequences of Paths for a Cyclic System (a Naturally Occurring Monocyclic Monoterpene), Including a Sum of Paths for Each Atom

atom	paths	sum of paths
1	1, 2, 2, 2, 3, 4, 1	15
2	3, 2, 2, 3, 4, 1	15
3	3, 4, 3, 4, 1	15
4	2, 3, 5, 4, 2, 2	18
5	2, 3, 4, 5, 3, 1	18
6	3, 3, 3, 5, 2	16
7	2, 2, 2, 3, 5, 2	16
8	1, 1, 2, 2, 3, 5, 2	16
9	2, 4, 5, 2, 2, 3	18
10	1, 2, 2, 2, 3, 4, 1	15

have the same path enumerations. An example of this is provided by atoms a and b (which are clearly nonequivalent) in the structure:



which both have the code: 3,4,2. Although the individual atom codes are not unique, the complete list of path codes for all atoms of a molecule appears to be unique to the structure.[5]

As has been discussed elsewhere, such codes are of use in searching for regularities in available molecular data[6] and in searching for substructures embedded in larger molecular skeletons.[7] In cyclic and polycyclic structures, enumeration of the paths becomes progressively more involved as the number of cycles increases, since there are now multiple paths between a pair of vertices. Thus, some care is necessary to avoid duplicate counts or the omission of some paths. It is well known that path tracing (the traveling salesman problem) requires an exponential increase in computation time as the size of the graph increases. The technique described in the paper requires tracing all paths starting from every atom; hence it may appear to be practical only for very small structures. We have prepared a program[8] and applied it to polycyclic structures having as many as 24 carbon atoms and six or seven rings and found the enumeration of all paths practical. Such polycyclic structures could have several thousands paths which are typically found in no more than few seconds of computation.

In the present context, we are interested in characterizing a molecule as a whole, rather than its individual atoms. However, the use of the list of all atom codes is cumbersome and not necessarily practical. For this reason, it seems desirable to explore some contractions of the complete list of atom codes in order to determine whether they will be adequate for our purpose. Two possibilities naturally suggest themselves: (1) atom path codes are truncated after a certain specified path length; (2) atom codes are combined in a single sequence of path numbers for a molecule. The most natural contraction is achieved by simply summing the contributions of all atoms for each path length. The first alternative may be of more interest when one wants to compare atomic properties within a molecule or properties among atoms of different molecules, since then one is primarily interested in the local atomic environment. Here, truncation can be justified on the grounds that long paths provide information on more distant atoms whose interactions necessarily have much less effect on local environment. For the second alternative, it is difficult a priori to appreciate how useful such molecular path sequences could be. Obviously, on contracting the information given by the list of atom path codes into a single molecular code, much of the initial structural information may be lost. Nevertheless, there are indications that such contracted codes may preserve important elements of the structure and remain useful descriptors of a system. The first entry in the derived molecular code gives the number of bonds; the second gives the number of adjacent pairs of bonds. In many atom and bond additive schemes, these make the dominant contributions.[9] So, the molecular codes described here could serve in studies of selected molecular additivity relations. In the case of acyclic structures, the derived codes are the same parameters that Platt[10] suggested as potentially useful quantities in studies of isomeric variations in alkanes. They also appear as coefficients in polynomials considered by Altenburg[11] for deriving the average radius of acyclic molecules. Therefore, it is not surprising that path enumeration and sequences of path numbers may reflect some structural features of a system. In the following discussion of similarity among structures (molecular skeletons as expressed by graphs), we will use the molecular path codes outlined above and show that they can indeed point to similarity.

SIMILARITY OF GRAPHS

It is plausible to expect that, starting with what intuitively appear to be similar structures, we may derive path sequences which will also be similar. In Figure 3, this is illustrated with a selection of acyclic, monocyclic, and bicyclic nine-atom molecular skeletons.

We augmented the path list with an initial entry giving the number of atoms (which can be formally viewed as paths of length zero). Under each structure, we show the single list of paths for all atoms in the molecule. Thus 9, 10, 8, 6, 5, 4, 3 under the skeleton of 2,2-dimethylheptane means that for this molecule there are 9 atoms (zero path lengths!), 10 bonds (paths of length one), 8 adjacent pairs of bonds (paths of length two), etc. All six nonanes in the figure have similar sequences. Furthermore, the sequences of cycloheptanes are of a similar form, differing from the sequences of cyclohexane structures or bicycloheptane derivatives. The differences between the sequences for molecules of the same class are clearly less pronounced than the differences for molecules belonging to different classes. We conclude that our intuitive notion that similarity among skeletal forms is reflected by similarity in path sequences is valid—at least for the sample of structures considered. The converse (i.e., that similar paths may point to similar structures) is less obvious. Two different objects can produce a similar projection. Hence, if one uses projections in which the information is reduced, it could be dangerous to draw conclusions about the similarity of objects in general. Nevertheless, we *assume* (as a working hypothesis) that molecules with similar path codes will also have similar structural features and can be expected to show some physical and chemical similarities. The situation using the suggested path codes is somewhat different from simpler cases of projections and the accompanying loss of information. Observe that here (1) the sequences of path numbers appear *unique* to the structure; (2) the entries of the codes are *constrained*, and not all sequences of numbers will correspond to structures. The codes have to be *legitimate*—to borrow language from graph reconstruction problems.[12] This then indicates that the *sequences of path numbers may preserve more information than is obvious*. In order to determine whether path codes based on enumeration of all paths of different length within a molecular skeleton still conserve important structural characteristics, we have applied the codes to a structural similarity problem.

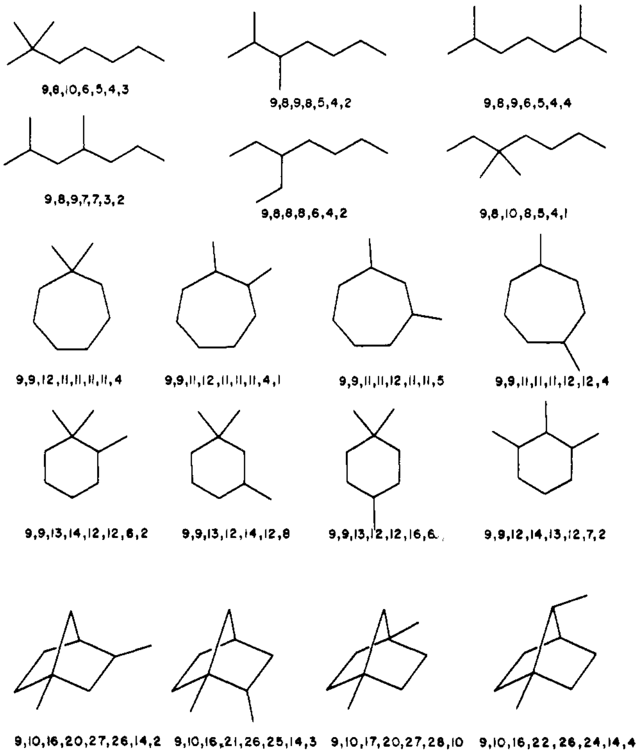


Figure 3. A selection of acyclic, monocyclic, and bicyclic nine atom systems which illustrate that apparent similarity of the structural forms is also reflected in the corresponding molecular codes based on enumeration of all paths of different length (summation of the corresponding atomic contributions).

For this purpose, *graphs* were considered rather than actual molecular models. This is not a restriction since detection of similarity among graphs is an even more general problem than that question of identifying similarity among rigid skeletal forms. Although an intuitive approach may help in comparing three-dimensional rigid structures, in the case of graphs it can be misleading. The *same* graph may be represented in very different forms, as illustrated in Figure 4 on the Desargues–Levi graph, introduced into chemistry by Balaban and co-workers in their study of isomerization of carbonium ions.[13] Various representations, taken from different sources,[14] show different aspects of the same connectivity and clearly show that the *appearance* of graphs is, in the present case, an irrelevant quality. One has to base similarity comparisons on some *graph invariants*, such as various subgraphs. Paths of different length provide such useful invariants, especially since preliminary work has suggested that generally for different graphs different sequences are derived.[15] Hence, molecular codes based upon enumeration of paths are adopted as a means of comparison of structures. When this is done, the problem of finding similarity among structures (graphs) now becomes the problem of finding similarity among sequences of integers. Such sequences can be compared and ordered as Muirhead demonstrated many years ago.[16] It is very natural to base similarity tests on these sequences. One can view such a sequence as a vector in n-dimensional Euclidean space (n being determined by the longest path or the truncation size) and call

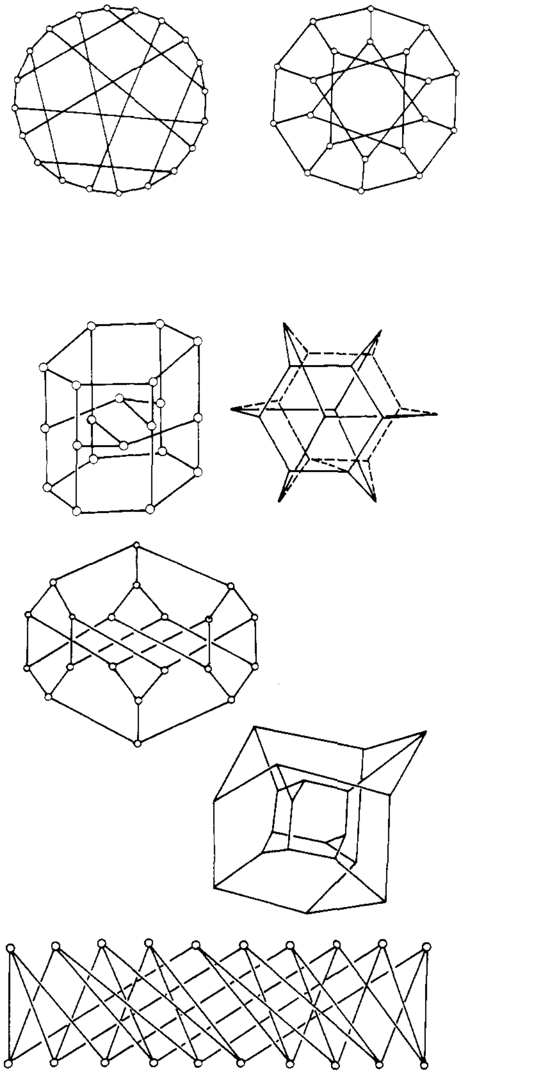


Figure 4. Various apparently different representations of the same graph (Desargues–Levi graph) illustrating that the appearance of a graph is irrelevant in characterization of graph similarity.

those vectors *similar* which lie in the same region of the space. As a measure of the degree of similarity, one can then take the distance between the points defined by two sequences:

$$D_{ab} = [\sum_i (a_i - b_i)^2]^{1/2}$$

where a,b refer to the path codes for different structures and the summation is carried over indices i which indicate the lengths of the paths. As an index of similarity, we use $S = 1/D$. Such a metric has some convenient properties. The maximal similarity is obtained for $S = 1$, which is the case when two structures differ only by a single entry in their path sequences. As the number of differences in the two sequences increases, S values decrease, first sharply, and then gradually approaching zero asymptotically for structures showing increasing differences in the number of paths of different length encountered. Singularity (i.e., $D = 0$ or $S = \propto$) would signify *identical* systems (isomorphs) and those rare instances when two molecules are isocodal, hence equivalent in the present similarity test.

ILLUSTRATION

In order to illustrate the use of molecular path numbers for establishing more rigorously (and certainly less arbitrarily) similarity among structures, we consider the molecular skeletons diagrammed in Figure 5. These skeletons were computer-generated by Smith and Carhart[17] in their study of mechanistic models which could explain formation of naturally occurring terpenes by head-to-tail linking of isoprene units. In Table III, molecular codes for the 29 monoterpenes of Figure 5 are listed. Table IV is only a part of a 29 × 29 (dis)similarity table in which D^2 rather than D are tabulated.[18] In the complete table, the values for D^2 lie in the range 3 to greater than 300, showing quite a spread of similarity values. It is instructive to look at the structures which the table suggests are very similar as well as those structures which are of limited similarity (i.e., indicated as structurally unrelated). Figure 6 contains all pairs found to show the greatest similarity among themselves (defined as $S \geq 1/\sqrt{5}$). An examination of the selected structures is encouraging. They differ only slightly, usually in the position of one substituted bond, the rings are of similar size, and the exocyclic attachments are

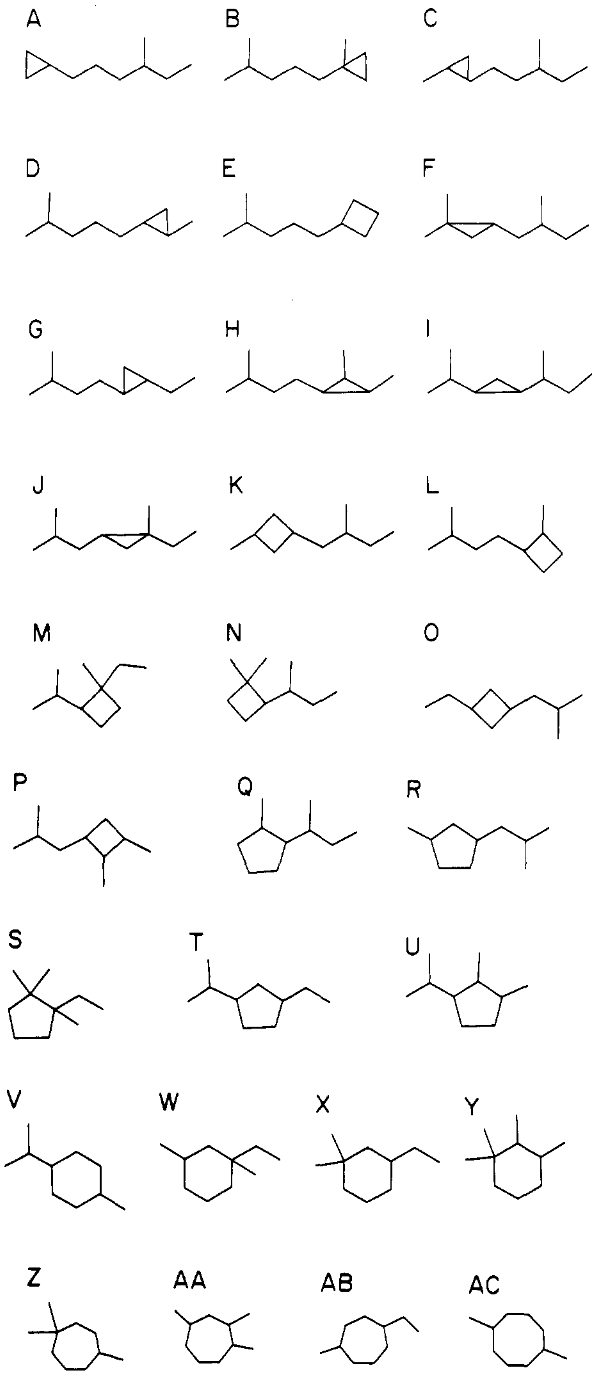


Figure 5. The set of 29 monocyclic monoterpenes used for illustration of the rigorous approach to finding similarities of structures and graphs.

Table III. Molecular Path Sequences for a Set of 29 Monocyclic Monoterpenes Generated by a Computer Program as Reported by Smith and Carhart[17]

structure	paths	sum of paths	structure	paths	sum of paths
A	10, 12, 10, 8, 7, 7, 6, 2	62	P	10, 14, 16, 15, 12, 9, 2	78
B	10, 14, 11, 8, 7, 8, 4	62	Q	10, 13, 16, 16, 11, 7, 4, 1	78
C	10, 13, 12, 9, 9, 9, 5, 1	68	R	10, 13, 13, 16, 12, 9, 6	79
D	10, 13, 11, 9, 8, 8, 7, 2	68	S	10, 16, 20, 17, 10, 5, 2	80
E	10, 12, 12, 9, 8, 8, 6, 4	69	T	10, 13, 15, 16, 14, 11, 2	81
F	10, 15, 14, 11, 12, 8, 2	72	U	10, 14, 17, 18, 13, 8, 2	82
G	10, 13, 12, 11, 10, 9, 5, 2	72	V	10, 13, 14, 14, 16, 12, 4	83
H	10, 14, 14, 11, 10, 10, 4	73	W	10, 14, 15, 16, 15, 10, 3	83
I	10, 14, 15, 15, 11, 7, 2	74	X	10, 14, 14, 16, 16, 10, 4	84
J	10, 15, 15, 13, 12, 7, 2	74	Y	10, 15, 17, 16, 14, 10, 3	85
K	10, 13, 14, 13, 10, 10, 4	74	Z	10, 14, 13, 13, 15, 15, 6	86
L	10, 13, 14, 11, 10, 9, 5, 2	74	AA	10, 13, 14, 14, 14, 14, 7, 1	87
M	10, 15, 19, 14, 10, 5, 2	75	AB	10, 12, 13, 13, 14, 15, 7, 4	88
N	10, 15, 18, 16, 10, 6, 2	77	AC	10, 12, 12, 12, 12, 14, 12, 4	88
O	10, 13, 14, 14, 12, 10, 4	77			

[a] Naturally occurring skeletons are: M, T, U, V, X, and Y (all 29 skeletons are shown in Figure 5).

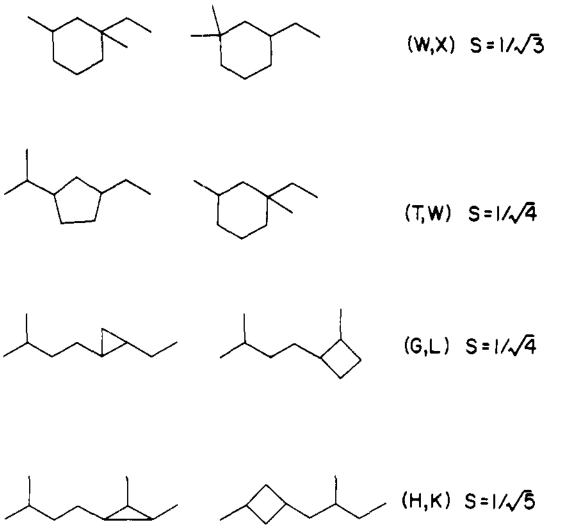


Figure 6. A few pairs of the most similar skeleton forms of Figure 5 with the degree of similarity indicated by the reciprocal distance of the corresponding coordinate points in the multidimensional vector space defined by the path lengths.

of similar length. Differences of one bond length or one ring atom increase or decrease appear to be the most that such a similarity test tolerates. Furthermore, it is seen that the structures suggested as dissimilar by Table III are indeed found to have large structural differences. Several cases are illustrated in Figure 7. These examples contain rings of much different size and have considerably different substituted exocyclic parts.

APPLICATIONS

As one possible application of the use of the proposed similarity measure, we suggest its use for selection from among a set of hypothetical monocyclic monoterpenes, those of potential interest in natural products chemistry. Smith and Carhart[17] have suggested the possibility that a novel skeletal form might be overlooked if the weight of arguments for its natural occurrence relies on biogenetic considerations alone. However, their attempt to identify a few potential novel skeletons was plagued by a proliferation of structures, most of which are not expected to be of particular interest. The difficulty lies, once the structures are available, in *selecting* those structures which may eventually be found in nature. Using path codes, we are in a position to make various selections employing the similarity test outlined, providing *standards* for comparison are available. An important observation can be made when examining the 29 × 29 table of

Table IV. A Portion of the Similarity Table for the 29 Structrues of Figure 5

.....									.....
....	S	T	U	V	W	X	Y	Z	.....
S		87	32	143	81	106	53	210	.....
T			19	14	4	11	10	47	.....
U				55	17	30	11	110	.....
V					12	9	26	17	.....
W						3	6	39	.....
X							15	40	.....
Y								61	.....
Z									.....
.....									.....

[a] The selected part illustrates the great similarity of structure W with several naturally occurring monoterpenes.

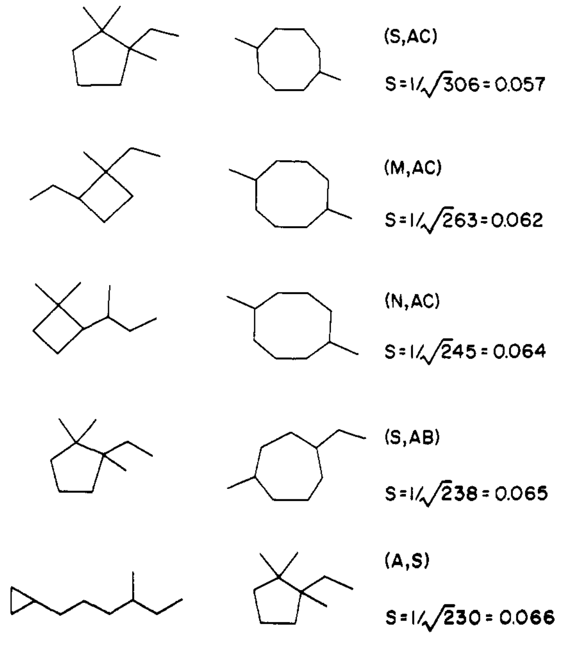


Figure 7. The pairs of skeletons of Figure 5 found to be the most dissimilar as measured by the differences in path numbers (expressed as a distance in multidimensional Euclidian space).

similarities of hypothetical and naturally occurring monocyclic monoterpenes. *The naturally occurring skeletal forms* (especially the five- and the six-membered ring systems T, U, V, X, and Y; see Table III) *are rather similar among themselves,*

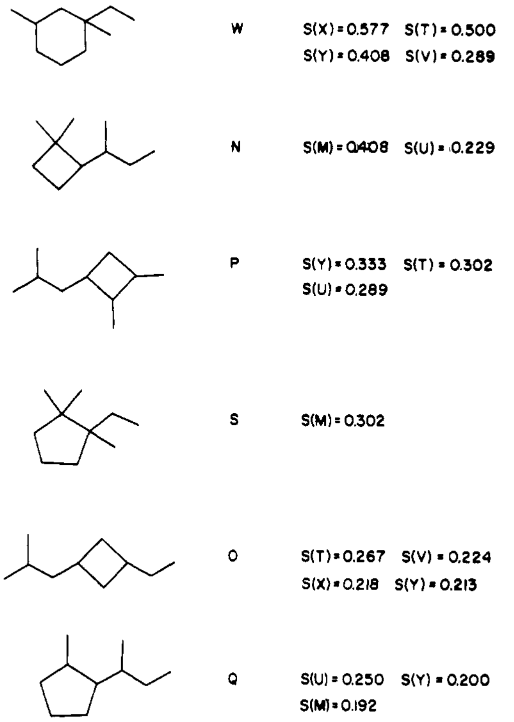


Figure 8. The structures most similar to the naturally occurring monocyclic monoterpenes. The degree of similarity with one or several common monoterpenes is indicated.

the similarity being based on enumeration and comparison of path numbers as described. So, we *postulate* that naturally occurring skeletons, even though the combinatorial possibilities are enormous, are similar among themselves. This postulate is plausible on the basis of chemical taxonomic arguments, which could suggest that mutations in (plant) species can gradually alter or modify some detail of terpene synthesis resulting in somewhat modified structures, while more drastic changes would have much less probability of survival.

In Figure 8, the monoterpenes which show the most similarity with the naturally occurring skeletons are listed. The naturally occurring forms are the same ones which were selected by Smith and Carhart on the basis of their appearance in texts on steroids.[19] For this comparison, we have chosen $S \geq 0.250$, but this can be changed according to needs. Among the structures of particular interest, appears structure W which *simultaneously* shows similarity with most of the already discovered naturally occurring skeletons. Therefore, it is suggested this structure is the prime candidate, among those considered, to be found in nature. Other skeletal forms in Figure 8 also should be considered as potential candidates since their degree of similarity to naturally occurring skeletons, for some, approaches (or even surpasses) the similarity found among the selected naturally occurring structures. The structure M, which is one of a naturally occurring skeleton, appears to show an appreciable degree of structural departure from the other natural monocyclic terpenes (five- and six-membered skeletons). It may be a member of another "family" of naturally occurring monoterpenes. It is especially interesting to find, among the structures of Figure 5, a few which are *both* fairly similar to structure M and structure U (which is a member of the majority group of naturally occurring compounds). These are Q and I, which can, thus, establish a "link" between the two groups showing larger differences.

In conclusion, if the outlined approach is found substantially correct in future work and is verified in other applications, it may provide important guidance to structural chemists. We have demonstrated here how the enumeration of paths in a molecular structure can help in restricting an impracticably large output of hypothetical structures (though the relatively small set of 29 monoterpenes taken for the example does not fully illustrate the potential power of the scheme). In addition to use for filtering many of the undesirable forms, molecular codes and the similarity test can be used in situations *when too few structures are generated* and yet one is interested in examining several additional structurally related forms. This, typically, is the situation when one applies pattern recognition or search techniques to find a structure which is compatible with available spectral information.[20] Here, once a single or few structures have been selected, one can use them as standards and search files for structures showing the greatest structural similarity with the selected standards. Work in this direction is underway. We should also add that the outlined scheme for finding similarity among structures is simply generalized (when appropriate) by introducing different weights for paths of different length. In an extreme case, the weighting corresponds to truncation of codes. Finally, in the illustration, we limited discussion to similarity among monocyclic monoterpenes. However, some of the skeletal forms of Figure 5 may show similarity with selected acyclic or bicyclic monoterpenes, though a change in the cyclization affects the number of paths considerably and the approach outlined may not be suitable for characterizations of similarities that occur upon ring closure.

ACKNOWLEDGMENT

One of (M.R.) would like to thank R. S. Hansen of Ames Laboratory for his support and encouragement. Partial support of this research by a National Science Foundation Grant CHE-76-21295 to the University of Nebraska is gratefully acknowledged. We also would like to thank P. C. Jurs and A. Zamora for helpful suggestions which improved the manuscript.

REFERENCES AND NOTES

(1) (a) Ames Laboratory; (b) University of Nebraska—Lincoln.

(2) F. Harary, "Graph Theory", Addison-Wesley, Reading Mass., 1969.

(3) Yu. A. Schreider, "Equality, Resemblance and Order", Mir Publishers, Moscow, 1975 (in English).

(4) L. J. Soltzberg and C. L. Wilkins, *J. Am. Chem. Soc.*, **99**, 439 (1977).

(5) This is a CONJECTURE, which has yet to be proved or disproved (by presenting a counter-example).

(6) M. Randić, *J. Chem. Soc., Faraday Trans. 2*, submitted for publication.

(7) M. Randić, *J. Chem. Inf. Comput. Sci.*, **18**, 101 (1978). M. Randić and C. L. Wilkins, *J. Chem. Inf. Comput. Sci.*, accompanying paper in this issue.

(8) A computer program (in PL/1, Fortran and Basic) is available. For more details see: M. Randić, G. M. Brissey, R. B. Spencer, and C. L. Wilkins, *Comput. Chem.*, in press.

(9) C. H. F. Hameka, *J. Chem. Phys.*, **34**, 1966 (1961); M. Randić, *Chem. Phys. Lett.*, **53**, 602 (1978).

(10) J. R. Platt, *J. Phys. Chem.*, **56**, 328 (1952).

(11) K. Altenburg, *Kolloid-Z.*, 178, 112 (1961). There is a trivial difference by a factor of 2, between Platt's and Altenburg's parameter.

(12) Reference 2, p 13.

(13) A. T. Balaban, D. Fărscasiu, and R. Bănică, *Rev. Roum. Chim.*, **11**, 1205 (1966).

(14) H. S. M. Coxeter, *Bull. Am. Math. Soc.*, **52**, (1946); A. T. Balaban, *Rev. Roum. Chim.*, **18**, 855 (1973); P. C. Lauterbur and F. Ramirez, *J. Am. Chem. Soc.*, **90**, 6722 (1968); K. E. DeBruin, K. Naumann, G. Zon, and K. Mislow, *ibid.*, **91**, 7031 (1969); K. Mislow, *Acc. Chem. Res.*, **3**, 321 (1970); M. Gielen and J. Nasielski, *Bull. Soc. Chim. Belg.*, **78**, 339 (1969); M. Randić, *Int. J. Quant. Chem.*, in press.

(15) The statement is based on examination of over 200 acyclic, cyclic, and polycyclic structures among which there are numerous isospectral graphs. So far only two cases of *isocodal* (i.e., having all path numbers equal) structures have been found: 2,4-dimethyl-4-ethylhexane–2,2-dimethyl-3-ethylhexane; and 2,3,4-trimethylhexane–3,3-methylethylhexane.

(16) R. F. Muirhead, *Proc. Edinburgh Math. Soc.*, **21**, 144 (1903); cf. G. H. Hardy, J. E. Littlewood, and G. P. Pōlya, "Inequalities", Cambridge University Press, London, 1934.

(17) D. H. Smith and R. E. Carhart, *Tetrahedron*, **32**, 2513 (1976).

(18) This is done for convenience: knowing that differences of 1, 2, 3, and so on in path numbers will introduce contributions of 1, 4, 9, and so on, respectively, D^2 can be visually partitioned into such contributions (for

small D values). One can also see that a single difference in value of path numbers by two introduces the same dissimilarity as four changes (increase of decrease) by a value of 1. The *metrics*, however, require that D rather than D^2 be used, since D satisfies the triangular rule: $D_{ab} \geq D_{ac} + D_{cd}$, a rule not necessarily obeyed by D^2.

(19) T. K. Devon and A. I. Scott, "Handbook of Naturally Occurring Compounds", Vol. II, "Terpenes", Academic Press, New York, 1972, pp 5-6.
(20) P. C. Jurs and T. L. Isenhour, "Chemical Applications of Pattern Recognition", Wiley, New York, 1975.

A Linked-Path Connection Table with Substructural Atom-Ordering

R. GEOFF. DROMEY*

Research School of Chemistry, Australian National University, Canberra, A.C.T. 2600, Australia

Received June 15, 1977

A new connection table formalism is introduced. The procedure for producing a canonical atom-ordering preserves molecular structural connectivity in a more explicit and direct way than existing systems. Consequently, substructure searching is made a much simpler process. The method focusses on atom connectivity and so judgments about bond characterization are avoided.

INTRODUCTION

Connection tables[1,2] achieve probably the simplest and most explicit topological description of molecular structure. They can be derived for all structures that are described in terms of specific atoms and bonds. In most connection table representations some form of atom-ordering procedure[3] is used to obtain a unique representation for each molecule. These simple algorithmic atom-ordering procedures invariably tend to breakup and inherent molecular "connectedness" that is conveyed so explicitly in the two-dimensional structure diagrams; that is, connectedness (in an atom adjacency sense) is sacrificed for a canonical representation. This means that substructure searching must always be on a strictly atom-by-atom basis. This complicates substructure searching of these systems.

In contrast, linear notations[4-7] tend to preserve the relationship between atoms and functionalities and so they make substructures much easier to identify. Consequently advanced pattern matching techniques that avoid the need to examine all characters in a representation may be employed.[8-10] Unfortunately, their rules for encoding have been far too complex to be used as a basis for canonicalization of connection tables. The valence-oriented rules used in the tree-structured linear notation described previously[7] depart from the formalism of earlier systems. They are sufficiently simple and algorithmic in nature to be used as a basis for producing a canonical connection table representation that to a large degree preserves molecular connectedness in much the same way as linear notations. The reason for adopting such an approach has been to produce a connection table representation that is usually more amenable to advanced pattern matching searches (in a way similar to line notations) than existing systems. It is also suitable for rapid visual interpretation. The unique numbering of atoms in the linear notation can provide the framework for the corresponding connection table ordering. In fact, on an atom-by-atom basis the two representations can be made directly compatible.

The four fundamental atom-ordering rules for the tree-structured linear notation are as follows.[7]

GENERAL RULES OF PRECEDENCE

At each stage in encoding a structure always choose to encode first

(A) that connected path with an atom of smallest valence attached earliest;

(B) where minimum valence does not resolve the path, choose to encode first the path with the atom of smallest atomic weight attached earliest;

(C) if resolution still has not been made encode along the path that contains an atom with the least number of atoms attached earliest;

(D) finally, if none of the other constraints resolves the path choose to encode along the path that has an atom with the least number of hydrogens attached earliest.

This set of rules is applied first in the sequence (A) to (D) wherever precedence must decide which atom or ring is to be encoded next.

A discussion of how the tree-structured formalism is used to obtain a unique atom numbering for a molecular structure is described in detail in the paper on the linear notation[7] and so will not be pursued further here. The present work will focus on the actual connection table representation assuming that a unique atom-numbering has been derived previously by applying the tree-structured precedence rules.

CONVENTIONAL CONNECTION TABLES

In order to allow for an adequate comparison to be made between the newly proposed formalism and existing systems, the conventions for the latter will be briefly reviewed. The most common set of descriptors used in connection table representations are element descriptor, "atom-connected-to", and order of bond forming the connection. Atoms are usually represented by their conventional atomic symbols while bonds are designated 1, 2, or 3 depending upon whether they are single, double, or triple bonds. To avoid citing each bond twice, only connections to lower numbered atoms are encoded. For ring systems this requires that ring-closure rank must be separately specified by taking into account atom numbers and bond codes and doing an ascending order sort.[3]

The conventional connection table representations for an acyclic and a ring compound are given in Tables Ia and IIa. Atom-ordering has been derived using Morgan's algorithm. The accompanying linked-path connection table representations for these structures will be compared and discussed in detail in the next section. Although structural features are represented explicitly in these conventional connection tables (Tables Ia and IIa), they suffer from a scattering effect imposed by the network formalism. The result is that a search for even simple functional groups becomes a tedious multistep procedure.[11]

* Address correspondence to author at Department of Computing Science, University of Wollongong, Wollongong, N.S.W. 2500, Australia.

0095-2338/79/1619-0037$01.00/0 © 1979 American Chemical Society