Computer Perception of Topological Symmetry

CRAIG A. SHELLEY and MORTON E. MUNK*

Department of Chemistry, Arizona State University, Tempe, Arizona 85281

Received November 14, 1976

A computer algorithm for the perception of equivalence classes from the topological representation of a molecular structure has been devised. The algorithm recognizes aromaticity in order to mimic closely the chemist's perception of topological symmetry. Application of the algorithm to the computer perception of stereochemical features in a molecular structure is discussed.

INTRODUCTION

Program CASE (Computer-Assisted Structure Elucidation) was conceived as a computer model of the intricate process by which the natural-products chemist reduces the chemical and spectroscopic data derived from an unknown compound to a molecular structure.[1] The manipulation and representation of a molecular structure required of the computer for this purpose can be facilitated by computer recognition of topological symmetry.[2] A new computer algorithm for the perception of topological symmetry in a molecular structure has been devised and incorporated into Program CASE.[3,4] The algorithm recognizes aromaticity in order to mimic the chemist's perception of topological symmetry. In this paper a detailed description of the algorithm is presented. Its application to the stereochemical representation of a molecular structure is also discussed. The algorithm also has proved useful in applications of ^{13}C NMR spectroscopy to chemical problems[5] and in the canonical representation and elaboration of molecular structures.[4,6]

For most compounds of carbon and related structures the experienced chemist easily recognizes topological symmetry upon inspection of the structural diagram without consideration of a set of instructions. On the other hand, computer perception of topological symmetry from the computer representation of a molecular structure does require a detailed set of instructions. For purposes of computer modeling, the process of perception by the chemist must be dissected. Although the method may vary with the chemist, one approach can be illustrated using as a simple example, 1-butanol.

1-Butanol has three immediately apparent classes of nonhydrogen atoms, methyl carbon, methylene carbon, and hydroxyl oxygen.[7] Further examination reveals that the methylene carbon atoms must be topologically nonequivalent since their neighboring atoms belong to different classes. One has hydroxyl oxygen and methylene carbon as nearest neighbors, another, two methylene carbons, and the third, a methyl and methylene carbon. Thus, the chemist recognizes five distinct topological classes of atoms. For more elaborate molecular structures, the chemist intuitively simplifies the process by utilizing more complex structural fragments as nearest neighbors; however, simple "atomic groups" as in the case of 1-butanol can be used in the analysis if the consequences of *all* neighbors, the nearest neighbors through the *n*th nearest neighbors, are considered.

For 1-pentanol, examination of nearest neighbors gives five classes of atoms rather than six, since two of the methylene carbons have identical nearest neighbors. An examination of the second nearest neighbors partitions these two methylene carbons into two distinct classes.

An algorithm for the canonical representation of connection tables based on this concept has been described by Ugi and co-workers.[8] The algorithm described in this paper embodies the concept in a manner which simplifies the computer implementation and perceives topological symmetry in molecular structures of any complexity and constitutional properties.

RELATED ALGORITHMS

The Morgan algorithm[9] partitions the atoms in a molecule into classes using "extended connectivity". Extended connectivity iteratively uses the connectivity of adjacent atoms to distinguish between atoms of the same class. Wipke[10] noted that extended connectivity is a measure of how centrally involved an atom is within a structure. The Morgan algorithm, as described by Wipke and Dyott,[10] is as follows: (1) set the extended connectivity (EC) of each atom to the number of nonhydrogen atoms to which it is bonded; (2) count the number of different EC values (NECV); (3) set the trial extended connectivity (TEC) of each atom to the sum of the EC values of adjacent atoms, unless the atom is primary, in which case its TEC is set at 1; (4) count the number of different TEC values (NTECV); (5) if NTECV is not greater than NECV, go to step 7; (6) set the EC value of each atom to its TEC value, set NECV to NTECV, and go to step 3; (7) done, the EC values are the final ones. Figure 1 traces the implementation of the algorithm for 1-butanol.

The efficiency of the Morgan algorithm can be increased if additional properties of the atoms are introduced, in particular, element type and the number of covalent (two-electron) bonds that join nonhydrogen atoms. These modifications serve to increase the initial partitioning of atoms into classes. For example, if the hydroxyl group of 1-butanol is initially distinguished from the methyl group, the methylene carbons are also partitioned into separate classes. In the iterative process, the previous class of an atom can provide further partitioning. This is accomplished by using the EC value of an atom as well as the neighboring atom EC values in calculating its TEC value. Weighting the previous class of an atom higher than the class of adjacent atoms also produces increased partitioning. We have modified the Morgan algorithm to embrace these concepts: (1) set the extended property (EP) of each atom equal to elemental type (e.g., C = 2, N = 3, O = 4, etc.) times ten plus the number of covalent (two-electron) bonds by which it joins to nonhydrogen atoms; (2) count the number of different EP values (NEPV); (3) set the trial extended property (TEP) of each atom to the sum of the EP values of adjacent atoms plus five times the EP value of this atom; (4) count the number of different TEP values (NTEPV); (5) if NTEPV is not greater than NEPV go to step 7; (6) set the EP value to its TEP value, set NEPV to NTEPV, and go to step 3; (7) done, the EP values are the final ones. Figure 2 traces this modified Morgan algorithm for 1-butanol.

This modified Morgan algorithm completely partitions the atoms into equivalence classes for most molecular structures. In our review of over 50 molecules of divergent structure type, only two molecules, **1** and **2** (Figure 5), failed to give correct results. In both cases, two constitutionally nonequivalent atoms had identical EP values upon algorithm termination. Partitioning is occasionally incomplete since the summing of

Sequential Order: 1 CH_3 – 2 CH_2 – 3 CH_2 – 4 CH_2 – 5 OH

Step No.	Assignment Performed	Atom No. 1	2	3	4	5
1	EC =	1	2	2	2	1
2	NECV = 2					
3	TEC =	1	3	4	3	1
4	NTECV = 3					
6	EC =	1	3	4	3	1
	NECV = 3					
3	TEC =	1	5	6	5	1
4	NTECV = 3					

Figure 1. Morgan algorithm applied to 1-butanol.

Sequential Order: 1 CH_3 – 2 CH_2 – 3 CH_2 – 4 CH_2 – 5 OH

Step No.	Assignment Performed	Atom No. 1	2	3	4	5
1	EP =	21	22	22	22	41
2	NEPV = 3					
3	TEP =	127	153	154	173	227
4	NTEPV = 5					
6	EP =	127	153	154	173	227
	NEPV = 5					
3	TEP =	788	1046	1096	1246	1308
4	NTEPV = 5					

Figure 2. Modified Morgan algorithm applied to 1-butanol.

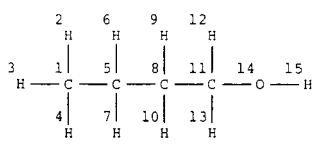
neighboring atom EP values is not necessarily equivalent to describing the set of neighboring atoms; i.e., chance may partition atoms into the same equivalence class even though they are topologically nonequivalent.

The Ugi algorithm[8] was designed for the canonization of a connection table, but may be utilized to identify topological symmetry. For this purpose the algorithm may be described as: (1) form equivalence classes of atoms of like atomic number, and arrange the classes in order of descending atomic number; (2) assign each member of an equivalence class the same tentative atomic sequence index (ASI) $k + 1$, where k = the number of atoms with higher atomic number; (3) select the lowest lying equivalence class with more than one member, if one exists, otherwise go to (8); (4) n gets 1; (5) assign a neighboring weight vector (NWV) for each atom consisting of the nth nearest neighbor ASI's arranged from left to right in ascending numerical order (the total number of entries in each NWV should equal the coordination number of the atom; therefore zeros are added as necessary); (6) new ASI's are assigned, with the atom bearing the lowest NWV receiving the lowest ASI available in the equivalence class; (7) if all atoms within this equivalence class have unique ASI's or if all neighbors have been considered, differentiation is complete; go to step 3, otherwise, n gets $n + 1$ and go to step 5; (8) done, the ASI's represent the topological equivalence classes. Figure 3 illustrates the application of the Ugi algorithm to 1-butanol.

TOPOLOGICAL SYMMETRY ALGORITHM

The topological symmetry algorithm possesses some of the characteristics of the three algorithms described above. However, it is highly efficient in computer implementation and is rigorous in its perception of topological symmetry. As with the Morgan algorithm, it extends atom environments throughout the graph, but includes the necessary atomic properties to produce complete partitioning.

Sequential Order:



Step No.	Assignments Performed		
1,2	Equivalence Classes		Tentative ASI
	(1) 0		1 (Final ASI)
	(2) ^{1}C, ^{5}C, ^{8}C, ^{11}C		2
	(3) H's		6
4	n = 1		
5,6	Atom	NWV	Intermediate ASI
	^{1}C	02060606	5
	^{5}C	02020606	3
	^{8}C	02020606	3
	^{11}C	01020606	2
7	n = 2		
5,6	Atom	NWV	Final ASI
	^{1}C	---	5
	^{5}C	020606060606	4
	^{8}C	010506060606	3
	^{11}C	---	2

Figure 3. Ugi algorithm applied to 1-butanol (hydrogen differentiation not considered).

The algorithm partitions the nonhydrogen atoms of a molecular structure into all possible equivalence classes by utilizing the class membership of nearest neighbors. The class membership of an atom is represented by an integer termed the class identifier. The algorithm can be described in terms of the following steps.

1. A class identifier (CI) is initially assigned to each nonhydrogen atom according to the number of bonds attached and element type. The CI value is a two-digit integer, the most significant of which specifies the number of covalent (two-electron) bonds joining the atom to nonhydrogen atoms, and the least significant of which designates atom type. The same integers for element type as described for the modified Morgan algorithm are used here, i.e., C = 2, N = 3, O = 4. Thus, a methyl group would be assigned a CI of 12.

2. Count the number of different CI values (NCI) and assign integers between one and NCI to the CI of each atom, where the smallest CI value becomes one and the largest CI value becomes NCI.

3. If NCI is equal to the total number of nonhydrogen atoms then go to step 7, else, assign a trial class identifier (TCI) to each atom. The format consists of five two-digit integers. The leftmost field contains the CI of the atom itself. The next four fields contain an ordered ascending list of nearest neighbor atoms CI's. The format of the TCI provides for up to four nearest neighbors. If there are less than four nearest neighbors the ordered ascending list is right justified in the four available fields and the leading digits are zero filled. The TCI describes the class membership of all nearest neighboring atoms and the previous class membership of the given atom.

4. Count the number of different TCI values (NTCI) and assign integers between one and NTCI to the TCI of each atom, where the smallest TCI value becomes one and the largest TCI value becomes NTCI.

5. If NTCI is not greater than NCI then go to step 7.

6. Set the CI of each atom to its TCI and set NCI to NTCI; go to step 3.

7. Done. The topological symmetry is represented by the class identifiers.

Sequential Order: $\overset{1}{CH_3}-\overset{2}{CH_2}-\overset{3}{CH_2}-\overset{4}{CH_2}-\overset{5}{OH}$

Step No.	Atom No.	Assignments Performed			
1,2,3,4		CI	CI (NCI=3)	TCI	TCI (NTCI=5)
	1	12	1	0100000003	1
	2	22	3	0300000103	3
	3	22	3	0300000303	5
	4	22	3	0300000203	4
	5	14	2	0200000003	2

6 CI(1) = 1, CI(2) = 3, CI(3) = 5, CI(4) = 4 and CI(5) = 2
NCI = 5 = number of nonhydrogen atoms in molecule

Figure 4. Topological symmetry algorithm applied to 1-butanol.

The algorithm as applied to 1-butanol is traced in Figure 4.

TREATMENT OF AROMATICITY

The topological symmetry algorithm was designed to recognize the special symmetry conferred on molecules possessing aromatic moieties while retaining the convenient Kekulé convention in designating such systems.

For unsaturated molecules, the "set" of nearest nonhydrogen neighbors can be defined in two ways. Neighboring atoms connected through multiple bonds may be considered as being in the "set" once, or, alternatively, by a number equal to the bond multiplicity. The first alternative is used in the modified Morgan algorithm. By way of illustration refer to structure **3** (Figure 5). Atom 2 may be considered as being connected to atom 1 either once or twice; i.e., the set of neighboring atoms is {1,3} or {1,1,3}, respectively.

For most structures, e.g., **3**, either implementation gives rise to the same end result; however, with aromatic systems the first method yields an answer consistent with the chemist's sense of symmetry, i.e., the chemical consequences of aromaticity. The second method of implementation treats the structure as actually drawn. Put differently, the first method of implementation treats the resonance hybrid; the second method, the actual resonance form drawn. Azulene provides a useful illustration. Whether drawn as Kekulé structure **4** or convention **5**, the chemist naturally determines six classes of topologically different atoms, rather than ten. Using the common convention **4**, implementation via the first alternative predicts six equivalence classes; the second approach predicts 10 classes.

Substituted benzene compounds present a similar situation. Operating in the first mode of implementation, the algorithm recognizes the aromaticity of the benzene ring and correctly predicts five classes of atoms for toluene (**6**). Interesting molecules for which the topological symmetry was correctly predicted are shown as **7**–**16**. It can be noted some of these structures provide a substantial challenge to the intuition of the chemist. In all cases, the number of equivalence classes shown takes aromaticity into account.

The algorithm, when operating in the "aromaticity" mode, does not at this time distinguish conjugated cyclic systems with aromatic character from those without aromatic character. Thus, the algorithm attributes "aromatic character" to cyclobutadiene and cyclooctatetraene, for example. The "nonaromaticity" mode of implementation provides correct answers in these cases.

PERCEPTION OF STEREOCHEMISTRY

Although the topological symmetry algorithm is currently receiving attention for other reasons (see Introduction), its application to the stereochemical representation of a molecular

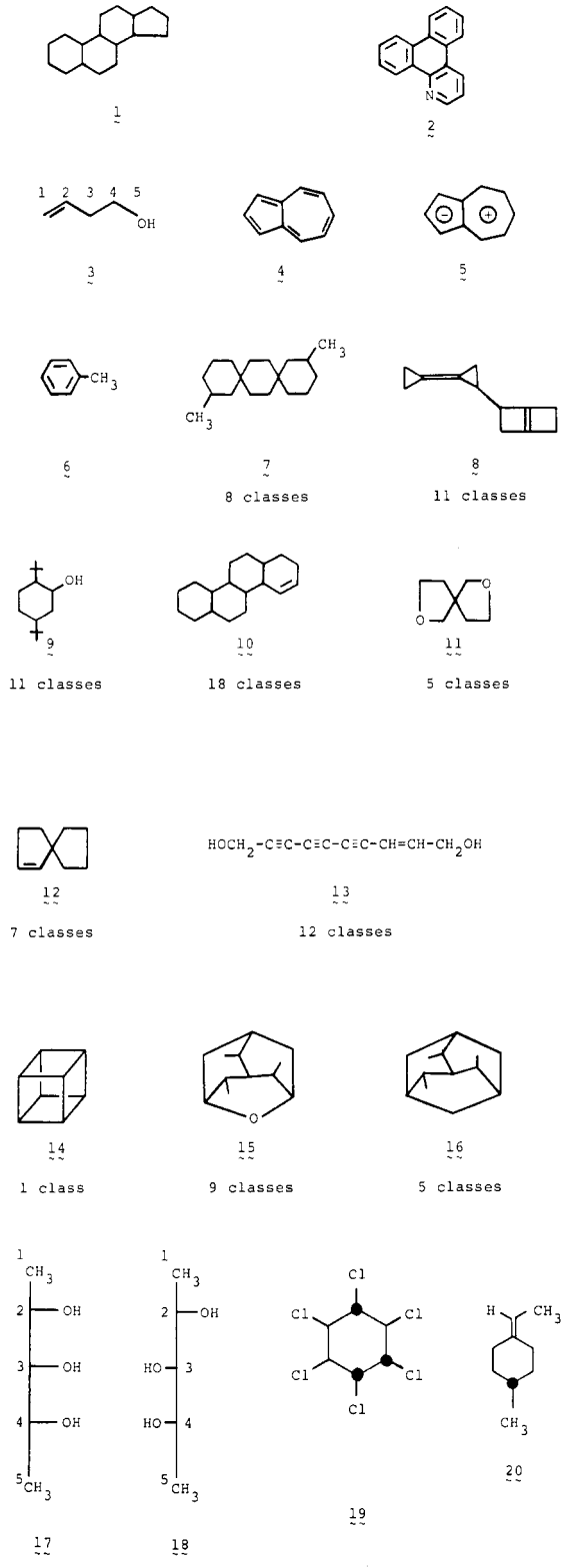


Figure 5. Pertinent molecular structures.

structure will be briefly considered. The algorithm is especially suited to this purpose because of the rigor of the assignment of atoms to equivalence classes.

Wipke and Dyott[10] have developed a stereochemically unique naming algorithm in which the connection table contains an ordered list of nearest neighbors for each atom bearing three or four nonhydrogen attachments and for each carbon–carbon double bond bearing no more than one hydrogen at each unsaturated site. The necessary configurational information is derived from the connection table by sequencing

the list of atoms with the Morgan algorithm and assigning "parity", even or odd, to each stereocenter.

The class identifiers derived from the topological symmetry algorithm are ideally suited for the assignment of parity to all stereocenters dependent on molecular topology. This perception of stereocenters is not rigorous since there are cases where equivalency may depend upon the chirality of attachments. New class identifiers are needed for a determination of configuration in these cases. For example, consider the diastereomers **17** and **18**. In structure **17** atom 3 is a stereocenter, but not in **18** since the chiralities of the two topologically equivalent attachments are identical. The perception of stereocenters dependent on the configuration of other stereocenters can be achieved by appending the parity property to the class identifier and reimplementing the topological symmetry algorithm.

The algorithm for stereocenter perception may be described as follows: (1) assign class identifiers using the topological symmetry algorithm; (2) use the class identifiers to define the parity of each stereocenter resulting from the molecular topology; (3) append the parity to the class identifier and again partition into classes; (4) reassign the parity of each stereocenter using the new class identifiers; (5) done. It should be noted that chiral structures without stereocenters derived from topological properties are not amenable to this approach. For example, chiral structures **19** and **20** possess no stereocenters that result from molecular topology and require special treatment.

CONCLUSIONS

A simple approach to computer perception of topological symmetry which attempts to mimic that of the chemist has been described. Recognition of the implications of aromaticity is built into the program. The algorithm may be extended to include stereochemical perception.

The algorithm has been used with a large number of examples, and we believe it to be rigorous in the identification of equivalence classes. It is amenable to computer implementation and is efficient and relatively fast in execution. At no time does the algorithm require the examination of atom sets removed by more than one bond from the atom under consideration. This speeds computation since nearest neighbors are explicitly represented by the connection table, whereas further removed atom sets require time for identification.

ACKNOWLEDGMENT

The authors are pleased to acknowledge the support of this project by the National Institutes of Health (GM 21703).

REFERENCES AND NOTES

(1) (a) D. B. Nelson and M. E. Munk, *J. Org. Chem.*, **35**, 3852 (1970); (b) F. J. Antosz, D. B. Nelson, D. L. Herald, Jr., and M. E. Munk, *J. Am. Chem. Soc.*, **92**, 4933 (1970); (c) Computer Program STR 3, Ph.D. Thesis, B. D. Cox, Arizona State University, 1973.
(2) The topological representation of molecular structure defines only atom connectivities and element type, and is independent of the three-dimensional nature of the molecule. The term topological symmetry as used in this paper refers only to the symmetry properties of the topological representation. The symmetry properties of the molecular structure differ from its topological symmetry.
(3) H. B. Woodruff, C. A. Shelley, and M. E. Munk, "Interactive Structure Elucidation", Third International Conference on Computers in Chemical Research, Education, and Technology, Caracas, Venezuela, July 26–30, 1976.
(4) C. A. Shelley and M. E. Munk, "Computer Elaboration of Molecular Structures", Abstracts, 172nd National Meeting of the American Chemical Society, San Francisco, Calif., Aug 1976.
(5) C. A. Shelley and M. E. Munk, "Computer Simulated Spectra. Cmr Peak Prediction", Abstracts, 173rd National Meeting of the American Chemical Society, New Orleans, La., March 1977.
(6) C. A. Shelley, H. B. Woodruff, and M. E. Munk, "Interactive Structure Elucidation", Abstracts, 173rd National Meeting of the American Chemical Society, New Orleans, La., March 1977.
(7) Hydrogen atoms are implicitly represented internally when the element type and connectivity of nonhydrogen atoms are defined by a connection table. Since the topological symmetry of hydrogen atoms in a molecule is generally of no interest, it is not treated by the algorithm.
(8) J. Blair, J. Gasteiger, C. Gillespie, P. D. Gillespie, and I. Ugi, *Tetrahedron*, **30**, 1845 (1974); J. Gasteiger, P. D. Gillespie, D. Marquarding, and I. Ugi, *Top. Cur. Chem.*, **48**, 1 (1974).
(9) H. L. Morgan, *J. Chem. Doc.*, **5**, 107 (1965).
(10) W. T. Wipke and T. M. Dyott, *J. Am. Chem. Soc.*, **96**, 4834 (1974).

Canonical Numbering and Constitutional Symmetry

CLEMENS JOCHUM and JOHANN GASTEIGER*

Institute of Organic Chemistry, Technical University, D-8000 München 2, West Germany

Received November 1, 1976

An algorithm is described which assigns numbers to the atoms of a molecular graph in a canonical manner. It is proven that then there is a one-to-one correspondence between the constitution of a molecule and the bond matrix thus numbered. There are no molecules which lead to ambiguities as encountered with other procedures. During the numbering process constitutionally equivalent atoms are recognized. The algorithm is simple enough to be applied directly without any computational support to not-too-complex structures. A computer program based on that algorithm has been implemented and results are given.

I. INTRODUCTION

Coding molecular structures for the manipulation by computers has become of increasing importance.[1] Documentation and information retrieval, computer-assisted synthetic design, and structure–activity correlations are some of the more prominent fields of application. For various purposes a nonunique representation of chemical constitution suffices. But for the storage and retrieval of chemical structures a unique description is necessary to avoid multiple storage of the same molecule and to ensure exact matching between query structure and the information to be retrieved.

The extensive work going on in our laboratory in developing computer programs for the solution of chemical problems[2] is based on a mathematical model of constitutional chemistry.[3] A constitutional formula is represented by a BE matrix[3] for which various internal representations are being tested. In our synthetic design program EROS (Elaboration of Reactions for Organic Synthesis), the storage and retrieval of synthetic intermediates are required. As we need access to each in-

* Address correspondence to this author.