An Algorithm for Computing the Automorphism Group of Organic Structures with Stereochemistry and a Measure of its Efficiency

Krishna K. Agarwal†

Department of Computer Science, Louisiana State University in Shreveport, Shreveport, Louisiana 71115

Received July 1, 1997

The notational algorithm in SYNCHEM2 deals with constructing a unique connection matrix for a molecule. In addition, it identifies constitutionally equivalent and stereochemically equivalent atoms within a molecule. For molecules with centers of asymmetry, it also identifies whether or not a molecule is chiral. This algorithm has recently been extended to produce all the maps of an organic structure to itself, with and without stereochemical considerations. In addition, an efficiency measure has been introduced for such algorithms. The algorithm was fairly efficient; for example, all the maps of buckminsterfullerane to itself were obtained under 225 s of execution time on a 75 MHz Pentium computer with an efficiency rating of 3.125%.

INTRODUCTION

Computer programs for a notational system for organic molecules have been used successfully for several years by SYNCHEM2,[1,2] an automatic synthesis planning system that can manipulate stereochemical organic molecules with no manual intervention. Given a non-canonical connection matrix, the program constructs a unique connection matrix for a molecule. As indicated earlier,[1] the program produces a separate parity vector for the stereochemical information of the molecule, identifies constitutionally equivalent atoms, identifies stereochemically equivalent atoms, identifies centers of asymmetry, and determines the chirality of the molecule.

The programs used for generating a unique connection matrix were originally written in PL/I.[3] The new program is fairly efficient compared with earlier works,[4−6] although the general problem remains NP-complete.[1] The program (rewritten in Pascal) is now capable of finding all automorphisms (maps of the molecule to itself) with and without stereochemical considerations. The program accepts as input an easy-to-write stereochemical descriptor of the molecule in the form of a non-canonical connection matrix.

Few researchers have attempted to find all automorphisms of molecules with their stereochemistry taken into account, but Laidboeur et. al.[7] have begun to deal with the three-dimensional (3-D) aspects of molecules. In addition, constitutionally equivalent atoms are identified, as are those that are stereochemically equivalent. For molecules with centers of asymmetry, the program identifies whether or not a molecule is chiral.

THE REPRESENTATION OF THE MOLECULE

The connection matrix for a molecule, called the Topological Structural Description (TSD) is used for all input and output. The occurrence of bond resonance in an organic molecule complicates the generation of canonical TSDs and has been ignored here. SYNCHEM2 uses a SLING[1] for input and output of molecules, and the TSD for the internal representation of the molecule. Because we are interested in obtaining all maps of a molecule to itself, it is simpler to use the TSD for input and output as well. Several examples of the TSD for input of a molecule are available.[1]

THE OUTPUT PRODUCED BY THE PROGRAM

Consider the molecule shown in Figure 1 for example. The main objective of the new program is to obtain the automorphisms of the molecules with and without stereochemical considerations. The older program used in SYNCHEM2 does not include the algorithm for producing these automorphisms. For the molecule shown in Figure 1, the automorphisms produced by the program are as follows (only the carbon atoms are listed):

without stereochemistry:

```
2 3 6 5 1 4

3 2 5 6 4 1

5 6 3 2 4 1

6 5 2 3 1 4
```

with stereochemistry:

```
2 3 6 5 1 4

3 2 5 6 4 1
```

Secondary output produced by the program follows next. SYNCHEM2 programs include the algorithms for producing all of the following output, which are included here for the sake of completeness.

For the example under discussion, the TSDs and parity vectors generated are shown in Table 1. Each parity vector is of length 6 because there are six carbon atoms in the molecule. In our example, the TSD shown in Table 1 is the canonical TSD and P1 is the canonical parity vector. Ignoring the stereochemistry about the atoms, we obtain the

† E-mail: kagarwal@pilot.lsus.edu.

S0095-2338(97)00053-X CCC: $15.00 © 1998 American Chemical Society
Published on Web 02/12/1998

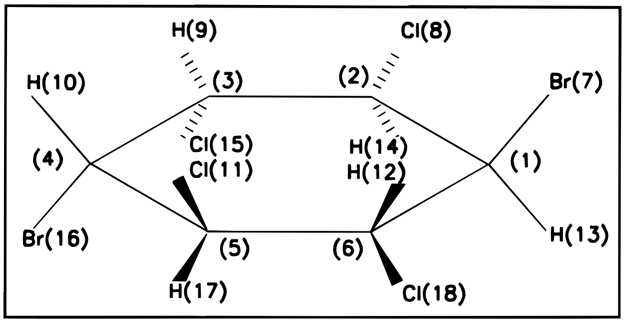

Figure 1.

Table 1. TSDs and Parity Vectors[a]

node-number	atom	up	down	left	right	in	out
1	C	5:1	2:1	Cl:1	H:1		
2	C	6:1	1:1	Cl:1	H:1		
3	C	5:1	4:1	Cl:1	H:1		
4	C	6:1	3:1	Cl:1	H:1		
5	C	3:1	1:1	Br:1	H:1		
6	C	4:1	2:1	Br:1	H:1		

[a] The two parity vectors are: P1:(−1,−1,−1,−1,−1,−1) and P2:(−1,−1,−1,−1,1,1).

Table 2. Constitutional and Stereochemical Equivalence Classes

node number	CE class	SE class
1	5	5
2	1	1
3	1	1
4	5	5
5	1	3
6	1	3

following constitutional equivalence (CE) classes for the molecule of Figure 1: {1,4}, {2,3,5,6}, {7,16}, {8,11,15,18}, {9,12,14,17}, and {10,13}. Atoms in distinct classes are not constitutionally equivalent. If the stereochemistry about the atoms is not ignored, we obtain the following stereochemical equivalence (SE) classes for the molecule: {1,4}, {2,3}, {5,6}, {7,16}, {8,15}, {9,14}, {10,13}, {11,18}, and {12,17}. Atoms in different classes are not stereochemically equivalent.

The CE and SE class numbers that are assigned by the program for the carbon atoms are shown in Table 2. For more precise definitions of CE and SE, refer to Agarwal,[3] Figueras,[5] and Rucker.[6] The program detects centers of asymmetry in a molecule. It also detects whether or not a molecule is chiral. The molecule shown in Figure 1 is determined to be chiral by the program.

SOME DETAILS OF THE ALGORITHM

The details of the algorithm are covered in Agarwal and Gelernter[1] and in Agarwal.[3] A very brief outline of the algorithm is given next to provide a context for the extensions that we have made to it.

The atoms of the input molecule are numbered arbitrarily to provide an input TSD for the program. The program renumbers the atoms in a manner that is independent of the original numbering of each atom to produce an invariant set of renumberings for the molecule. The set of numberings produced are as few as possible so that the algorithm is as efficient as possible. Each time the molecule is renumbered, a new TSD is constructed. Furthermore, each row of the new TSD is rearranged so that the list of neighbors appears in descending order. In doing so, only the constitution of the molecule is retained in a TSD and the stereochemistry of the molecule is isolated into a parity vector. Parity values for a tetrahedral carbon are either −1 or +1 and are −2 and +2 for olefin carbons. The "smallest" TSD generated for all the numberings is selected as the canonical TSD for the molecule. The smallest parity vector generated for all canonical TSDs is selected as the canonical parity vector. The canonical TSD produced by the program is distinct for molecules that are constitutionally distinct. The same TSD is generated for molecules that are constitutionally identical but stereochemically distinct. However, the canonical parity vectors generated are different.

The algorithm identifies constitutionally equivalent atoms as those that lead to the same smallest TSD during the renumbering process. Stereochemically equivalent atoms are those that not only lead to the smallest TSD but also to the smallest parity vector. The asymmetry of each atom is then determined and finally, the chirality of the molecule is determined.

The extension to the algorithm that deals with finding all the automorphisms is as follows. For each renumbering of the molecule if the TSD produced is identical to an older TSD, the renumbering is saved as a constitutional automorphism. If the parity vector produced is identical to the smallest saved parity vector, the renumbering is saved as a

Table 3. Summary of Results of Program

molecule name	# of maps attempted	# of maps without stereochemistry	# of maps with stereochemistry	chiral	time (s)	efficiency (%)
cyclohexane of Figure 1	4	4	2	yes	0.060	100
cyclohexane of Figure 1 with atoms 10 and 16 exchanged	4	4	2	yes	0.050	100
cubane	48	48	24	no	0.330	100
icosahedron of P atoms	120	120	—	—	0.880	100
dodecahedron of C atoms	480	120	60	no	5.170	25
buckminsterfullerane with one I atom	64	2	1	no	5.600	3.125
buckminsterfullerane with two I atoms opposite to each other	32	4	2	no	2.190	12.5
buckminsterfullerane	3840	120	60	no	223.110	3.125

stereochemical automorphism. However, if a new smaller TSD is produced, the process of saving automorphisms starts over.

A measure for the efficiency of this and other similar algorithms is defined as the percentage of constitutional automorphisms present in the molecule with respect to the total renumberings generated by the algorithm.

RESULTS AND CONCLUSIONS

The results obtained by running the program on a 75 MHz Pentium computer for a large variety of molecules are summarized in Table 3. It is interesting to note that for buckminsterfullerane, the total time consumed for finding all automorphisms with and without stereochemistry was only 223.110 s, a feat not achieved by any other program that the author is familiar with. Of the 3840 renumberings attempted, 120 led to constitutional automorphisms (thus leading to an efficiency measure of 3.125%). There were 60 stereochemical automorphisms.

In summary, the original algorithm provided us with the following information for the molecule: (1) the canonical TSD; (2) the canonical parity vector; (3) the set of parity vectors accompanying the canonical TSD; (4) the CE class numbers for each atom in the canonical TSD; (5) the SE class numbers for each atom in the canonical TSD; (6) the asymmetry information for each carbon atom in the canonical TSD; and (7) the information about its chirality. The extended algorithm also provides us with (8) the constitutional automorphisms of the molecule and (9) the stereochemical automorphisms of the molecule.

Few researchers have attempted to develop algorithms that can discover all automorphisms of molecules including their stereochemistry. Indeed, the statement at the end of Balasubramanian's paper "It appears at present that there cannot be a single technique that is uniformly elegant and efficient for all graphs"[4] is perhaps disproved by our program.

The program runs under Turbo Pascal 7.0 and is freely available for experimentation by sending an E-mail message to the author at kagarwal@pilot.lsus.edu. However, the author assumes no liability for its operation or maintenance (i.e., bugs in the program will not be fixed) or incorrect results, if any, that are obtained when the program is run. The source code supplied could be extended to handle resonant bonds.

REFERENCES AND NOTES

(1) Agarwal, K. K.; Gelernter, H. L. A Computer-Oriented Linear Canonical Notational System for the Representation of Organic Structures with Stereochemistry. *J. Chem. Inf. Comput. Sci.* **1994**, *3*, 463−479.

(2) Gelernter, H. L., et. al.: Empirical Explorations with SYNCHEM2. *Science* September 7, 1977, *197*, 1041−1049.

(3) Agarwal, K. K. Ph.D. Thesis, Department of Computer Science, State University of New York at Stony Brook, 1976.

(4) Balasubramanian, K. Computational Techniques for the Automorphism Groups of Graphs. *J. Chem. Inf. Comput. Sci.* **1994**, *3*, 621−626.

(5) Figueras, I Automorphism and Equivalence Classes. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 153−157.

(6) Rucker, G.; Rucker, C: Computer Perception of Constitutional (Topological) Symmetry. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 187−191.

(7) Laidbouer, T., et. al. Determination of Topo-Geometrical Equivalence Classes of Atoms. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 87−91.

CI9700539