Erroneous Claims Concerning the Perception of Topological Symmetry

RAYMOND E. CARHART

Machine Intelligence Research Unit, University of Edinburgh, Edinburgh, EH8 9NW Scotland,[1] and
Computer Science Department, Stanford University, Stanford, California 94305

Received September 7, 1977

Counterexamples are provided disproving two independent claims that a simple but accurate method had been found to compute classes of symmetrically equivalent atoms in a molecule. Both methods are good approximations, but are nontheless ad hoc techniques which can sometimes fail to discriminate between atoms which are, in fact, symmetrically distinct.

A recent issue of this journal carried two papers[2,3] dealing directly or indirectly with "inexpensive" (in the sense of low computational effort) methods of perceiving molecular symmetry.[4] More precisely, each article gives a set of rules for scoring[2] or comparing[3] atoms in a molecule with the claim that if the scores are equal, or if the comparison shows no difference, then the atoms are symmetrically equivalent (i.e., can be interchanged by some symmetry operation on the molecule). Though both methods are doubtless very good approximations in the sense that they almost always yield the classes of symmetrically equivalent atoms for typical chemical molecules, it is demonstrated below that neither is fully correct. There exist chemical graphs for which the number of distinct symmetry types of atoms is greater than that computed by the suggested algorithms. If these approximate methods are taken as accurate and used, as indicated by Shelley and Munk,[2] ". . . in applications of ^{13}C NMR spectroscopy to chemical problems and in the canonical representation and elaboration of molecular structures. . .", then it must be accepted that in some instances the wrong number of ^{13}C NMR peaks may be predicted, nonunique "canonical" numberings may be created, and correct structures may be overlooked during elaboration.

In developing programs of this kind, it is important for one to support each new algorithm for manipulating chemical structures with a solid foundation of graph-theoretic understanding, lest unexpected and perhaps unnoticeable errors occur in the output. Verifying that an algorithm is never *known* to fail when its output is compared with selected manually generated cases, or with the output of other programs based on totally different principles, does not constitute proof that the algorithm is valid beyond the specific cases which are tested. Such a "proof" leaves open the important question of the scope and limitations of the algorithm: How far beyond those test cases would one have to look to find one which fails? For this reason, my co-workers and I have, wherever possible, provided rigorous graph-theoretic proofs supporting the methods used in the DENDRAL programs.[5,6] It is crucial that readers of this journal understand the approximate nature of the cited symmetry-perception algorithms before applying the methods to their own problems.

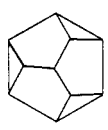
COUNTEREXAMPLES

Shelley and Munk[2] present an algorithm which is similar in concept to the atom-classification scheme presented by Morgan[7,8] as the initial stage of his algorithm for finding canonical (i.e., standard) numberings for molecular graphs. Shelley and Munk extend Morgan's scheme significantly, but contrary to their claim the improvement is not sufficient to guarantee accurate perception of topological symmetry in all cases. The initial step in either approach is to associate a score of some kind with each atom, representing its local characteristics (e.g., atom type, number of attached hydrogens) in the molecule. The rest of the algorithm consists of an iteration loop, at each stage of which a new score is computed for each atom based upon some function of the current score of that atom and the current scores of its immediate neighbors. The term "score" here is used rather loosely; in the case of Morgan's algorithm it is an integer while in that of Shelley and Munk's it is a five-element vector. The important point is that a score is an entity which is associated with an atom and which can be compared with the scores of other atoms in such a way that a strict "greater-less-equal" relationship can be defined. The iterations proceed until the number of distinct (i.e., non-equal) scores ceases to increase.

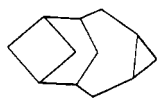
Without going into the algorithmic details, it is not difficult to show that symmetrically distinct atoms in some structures can give the same final score, quite independently of the method of score computation, as long as the general procedure outlined above is followed. Let an *initial class* of atoms in a molecule be a set of atoms which all have the same initial score. Furthermore, let two atoms i and j be *locally isomorphic* if they are in the same initial class and if the neighbors of i can be placed in a one-to-one correspondence with the neighbors of j such that the following holds true: If atom k neighbors i and atom l is the neighbor of j corresponding to k, then k and l are in the same initial class and the k–i bond order equals the l–j bond order. In less precise but more pictorial terms, two atoms are locally isomorphic if they and their nearest neighbors form superimposable substructures, not counting bonds interconnecting the neighbors.

Now if, for every initial class in a molecule, each atom in the class is locally isomorphic to every other atom in the class, then the general procedure presented above can never further partition the initial classes. That is, at every level of iteration, all members of an initial class must have the same score whether or not they are actually equivalent symmetrically. This is true because, at each iteration, a new score is computed using only the scores of the immediate neighbors of an atom. Because all atoms in each initial class are locally isomorphic, their first-level neighborhoods are identical and the scoring function will treat them identically. No discrimination will occur during the first iteration nor, by induction, in any subsequent iteration.

The remaining question, then, is whether molecules can be found which satisfy the above conditions for local isomorphism yet in which the initial classes are not identical with the true symmetry-equivalence classes. This is certainly true. Any molecule composed of trivalent groups CH and free of multiple bonds satisfies the condition of local isomorphism around every atom, and there are such molecules which have more than one symmetry-type of atom. Structure I will suffice as the



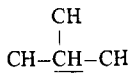
I



II

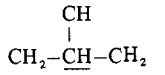
0095-2338/78/1618-0108$01.00/0

counterexample. The local environment of each atom is the following:

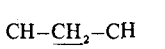


(Recall that bonds interconnecting the neighbors need not be considered for local isomorphism.) There is only one initial class, so the algorithm of Shelley and Munk can only recognize one final class. There are, though, three distinct classes of symmetrically equivalent methine carbons.

A more complex example is structure II in which each methine carbons has the local environment

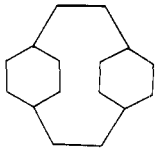


and each methylene carbon has the local environment

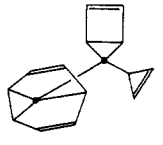


There are, in structure II, four nonequivalent classes of the CH_2 groups and three of the CH groups, but the approximate algorithm will yield only one class of each.

For molecules of this type, the resolution of nonequivalent atoms lying in the same initial class can only be accomplished by either including in the initial scores terms representing additional features of the local topology or by defining a function which, during each iteration, computes a new score based on more than just the first neighbors of an atom. The trouble with such extensions is that one never knows when some other refinement will be necessary. For example, one could (at some expense to the efficiency of the algorithm) modify the initial score by adding a quantity representing the number of three-membered rings[9] in which the atom participates (this is an easily computed feature of local topology) and extend the scoring scheme so that it accounts for the scores of not only each first, but also each second nearest neighbor. Either of these additions would solve the difficulty for both structures I and II, but not for structure III in which there

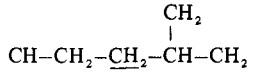


III

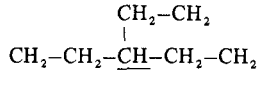


IV

are no three-membered rings (nor any smaller than six-membered, for that matter), in which each methylene carbon has the two-level neighborhood



and in which each methine carbon has the two-level neighborhood



The point is that the method is basically ad hoc, and, though it may be made correct for most known molecules, I do not believe it can be proven reliable until either the initial scores or the scoring scheme or both take into account the full topological context (i.e., the whole surrounding molecule) of each atom. But such an extension probably would make the algorithm equivalent to, and just as time-consuming as, formally correct algorithms which determine the actual symmetry group of a molecule and extract the equivalence classes from that.

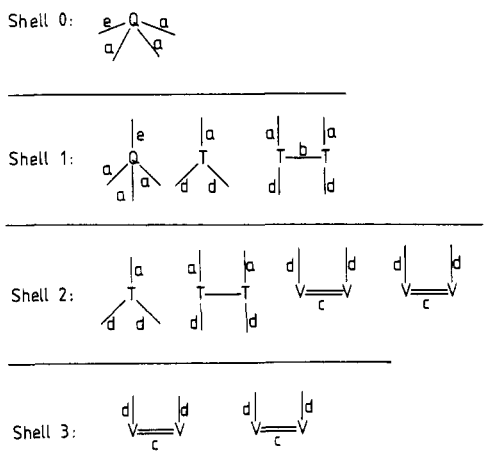


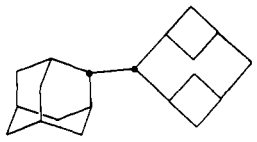
Figure 1. The "shell dissection" of structure IV, relative to either of the two heavily dotted atoms. The atom types Q, T, and V, as well as the bond types a–e, are defined in the text.

Jochum and Gasteiger[3] make the same error of letting a set of ad hoc rules substitute for true symmetry perception in their description of a scheme for generating canonical numberings for chemical structures. Central to their algorithm is their definition of "constitutional equivalence" between atoms, which is claimed implicitly to be the same as symmetry equivalence. Structure IV is an example of a molecule (perhaps not a stable one, but the point here is topological rather than chemical) in which two atoms, marked with heavy dots, are "constitutionally equivalent" without being symmetrically related.

Inspection shows that the marked atoms are not symmetrically equivalent. To show their "constitutional equivalence" one needs to consider three main features which are used in Jochum and Gasteiger's definition. First, the nonhydrogen atoms are classified according to type. The symbols Q, T, and V will be used to represent quaternary C, aliphatic CH, and vinylic CH atoms, respectively. Second, bonds are classified according to type as well, with bond order and the types of the terminal atoms distinguishing different bond types. The symbols a, b, c, d, and e will be used to represent bonds of type Q—T, T—T, V=V, V—T, and Q—Q, respectively. The third feature is the set of "shells" (neighbors spheres[3]) surrounding two atoms which are to be compared. The comparison consists of first matching atom and bond types for the zeroth shell (i.e., the atoms themselves), then testing whether the first shells match, then the second, and so on.

The term "match" here can be interpreted in a variety of ways. Jochum and Gasteiger have actually described a rather loose criterion under which shells are considered to be equivalent (that each atom in one shell must have a corresponding atom in the other shell which has the same atom type and the same numbers and types of bonds extending into the next outer shell), but one can imagine stronger tests, up to checking for actual superimposability (isomorphism) of corresponding shells. Even under this stronger test the two Q atoms of structure IV would be "constitutionally equivalent" because their shells are isomorphic at all levels. In Figure 1 is shown the "shell dissection" of the molecule relative to either of the two marked atoms in structure IV. Important topological information has been lost in cutting the shells apart and treating them independently, and thus these two atoms are erroneously taken to be equivalent. Other more chemically plausible counterexamples can be found, for example, structure V, in which the distinct dotted atoms are again equivalent according to Jochum and Gasteiger's definition.

If one tried to canonicalize structure IV according to Jochum and Gasteiger's method, one would have to make an arbitrary decision as to which Q would receive the very first



V

number. If the Q's really were equivalent, the choice would be immaterial, but as they are not, one may define two different "unique" numberings based on this initial arbitrary decision. The "proof" of the algorithm offered by Jochum and Gasteiger contains the implicit assumption that their ad hoc method of symmetry perception is accurate. If this were true, I believe their canonicalization algorithm could be proven correct, though a more detailed and precise chain of reasoning would need to be given.

CONCLUSIONS

The algorithm of Shelley and Munk fails in some cases because it does not consider the complete topological environment (i.e., the whole molecule) of each atom in computing scores. The algorithm of Jochum and Gasteiger does consider the whole molecule as "viewed" from a given atom, but in cutting the molecule into shells about that atom, it discards important connectivity information which may be needed to distinguish between nonequivalent atoms. It is my belief that a provably reliable algorithm for identifying symmetrically equivalent atoms must either explicitly compute the total symmetry group of the molecule or carry out a full atom-by-atom, bond-by-bond comparison of the total topological environments of atoms being compared.

ACKNOWLEDGMENT

I wish to thank both the Science Research Council and the National Institutes of Health (RR 00612) for their support of my current research.

REFERENCES AND NOTES

(1) Author's address until May 1, 1978.
(2) C. A. Shelley and M. E. Munk, *J. Chem. Inf. Comput. Sci.*, **17**, 110 (1977).
(3) C. Jochum and J. Gasteiger, *J. Chem. Inf. Comput. Sci.*, **17**, 113(1977).
(4) The term "symmetry" is used throughout this paper to refer to the topological symmetry of chemical graphs, rather than the geometrical symmetry of three-dimensional molecules. Topological symmetry is derived from the connectivity between atoms without reference to their spatial relationships.
(5) D. H. Smith and R. E. Carhart, *Tetrahedron*, **32**, 2513(1976), and previous papers in the series.
(6) R. E. Carhart, D. H. Smith, H. Brown, and N. S. Sridharan, *J. Chem. Inf. Comput. Sci.*, **15**, 124(1975).
(7) H. L. Morgan, *J. Chem. Doc.*, **5**, 107(1965).
(8) The accuracy of the Morgan canonicalization scheme is not in question here; it uses the atom-scoring algorithm in a theoretically sound manner to limit the number of potentially canonical representations which need be considered for a given structure. Rather than assuming that atoms with the same final score are symmetrically equivalent, Morgan's approach uses the fact that atoms with *distinct* final scores *cannot* be symmetrically equivalent.
(9) C. A. Shelley, private communication.

Computer Design of Synthesis in Phosphorus Chemistry: Automatic Treatment of Stereochemistry

F. CHOPLIN,* R. MARC, and G. KAUFMANN

Laboratoire de Modèles Informatiques appliqués à la Synthèse (ERA 671) ULP, Institut Le Bel, 67000 Strasbourg, France

W. T. WIPKE*

Department of Chemistry, University of California, Santa Cruz, California 95064

Received January 9, 1978

Computer treatment of the stereochemistry of phosphorus compounds is studied in the aim of building a program for the design of synthesis involving this type of chemistry. A naming algorithm has been developed to differentiate the isomers of pentacoordinated species. Techniques have been developed to conveniently and unambiguously describe the stereochemical changes which may occur at a pentacoordinated species during a reaction. Relative stabilities and possible intramolecular rearrangements between stereoisomers are also considered.

INTRODUCTION

The last several years have witnessed very rapid and outstanding progress in the field of computer design of synthesis. All previously reported work has been confined exclusively to the field of organic synthesis.[1a-i] However, some other areas of synthetic chemistry, such as those relevant to organophosphorus compounds, exhibit features which would expectedly permit their treatment by currently employed computer programs. Indeed, the theoretical and practical importance of these compounds makes application of computer-assisted synthesis techniques desirable. The fact that their synthesis can be designed with key steps involving general reactions makes a computer treatment feasible. If a program designed for organic synthesis could accomodate organophosphorus compounds, it would demonstrate the versatility of such a program. Moreover, many of the problems involved in organophosphorus chemistry are different from those encountered in organic chemistry; thus algorithms and strategies designed to solve these new problems will lead to a broader approach of the whole field. As a part of a project aimed at the development of the SECS program[2] for the design of organophosphorus synthesis, this paper describes the computer treatment of phosphorus stereochemistry. New problems arise owing to the different possible configurations about this atom, and to intramolecular rearrangements, which reduce the number of stable isomers. We shall describe first the un-

0095-2338/78/1618-0110$01.00/0