natural orbitals are the molecular orbitals and the occupation numbers are 0, 1, or 2. An index of the adequacy of a molecular orbital description is the closeness of the n_u's to integer values. (b) See, for example, H. F. Schaefer, III, "The Electronic Structure of Atoms and Molecules", Addison-Wesley, Reading, Mass., 1972.

(10) J. F. Arnett, G. Newkome, W. L. Mattice, and S. P. McGlynn, *J. Am. Chem. Soc.*, **96**, 4385 (1974).

(11) J. M. Leclercq, C. Mijourle, and P. Yvan, *J. Chem. Phys.*, **64**, 1464 (1976).

(12) J. Del Bene and H. H. Jaffé, *J. Chem. Phys.*, **48**, 1807 (1968).

(13) C. C. J. Roothaan, *Rev. Mod. Phys.*, **32**, 179 (1960).

(14) J. R. Swenson and R. Hoffman, *Helv. Chim. Acta*, **53**, 2331 (1970).

(15) J. A. Pople, *Acc. Chem. Res.*, **3**, 217 (1970).

(16) (a) P. J. Wagner, M. J. May, A. Haug, and D. R. Graber, *J. Am. Chem. Soc.*, **92**, 5269 (1970); (b) P. J. Wagner, A. E. Kemppainen, and H. N. Schott, *ibid.*, **95**, 5604 (1973); (c) E. Migirdicyan, *Chem. Phys. Lett.*, **12**, 473 (1972); (d) M. E. Long and E. C. Lim, *ibid.*, **20**, 413 (1973); (e) R. O. Loutfy and J. M. Morris, *ibid.*, **22**, 584 (1973).

(17) G. Marsh, D. R. Kearns, and K. Schaffner, *Helv. Chim. Acta*, **51**, 1890 (1968).

(18) A. Devaquet, *J. Am. Chem. Soc.*, **94**, 5160 (1972).

(19) (a) E. Drent, R. P. vander Werf, and J. Kommandeur, *J. Chem. Phys.*, **59**, 2061 (1973); (b) K. Kaya, W. R. Harshberger, and M. B. Robin, *ibid.*, **60**, 4231 (1974).

(20) C. E. Dykstra, R. R. Lucchese, and H. F. Schaefer III, *J. Chem. Phys.*, **67**, 2422 (1977).

Constitutional Symmetry and Unique Descriptors of Molecules

Wolfgang Schubert and Ivar Ugi*

Contribution from the Organisch-Chemisches Institut, Technische Universität München, 8000 München 2, West Germany.

Received January 3, 1977

Abstract: In systems for computer-aided synthesis planning the number of redundant synthetic intermediates generated can be greatly reduced by detecting the constitutional symmetry of molecules to be processed. The chemical shift pattern of NMR spectra yields information about constitutional symmetry with respect to the observed nuclei, and constitutional symmetry is important for the interpretation of NMR spectra. An algorithm has been designed which detects constitutional symmetry by finding the atoms which are similar with respect to automorphisms of the labeled graph underlying the structural formula of a compound. The information about the constitutional equivalence of atoms is then used to canonically order the atoms of the molecule. A canonical order is necessary for treating the stereochemical features of molecules. Furthermore, the canonical order of atoms directly leads to a unique representation of molecules. This unique representation is necessary to identify molecules when generating and evaluating the reaction paths in systems for automated synthetic design. Since identifying molecules is based upon the results of the algorithm for finding constitutional symmetry the overall efficiency of such systems can be substantially increased.

Introduction

The chemical constitution of a molecule, or an ensemble of molecules, is determined by the number and kind of atoms which it contains and those pairs of neighboring atoms which are connected by covalent bonds.

A chemical constitution is usually described by a constitutional formula. It consists of atomic symbols which are connected by lines. The atomic symbols represent atomic cores. They consist of the atomic nuclei and the inner electrons. The lines in constitutional formulas refer to covalent bonds which correspond to valence electrons in orbitals belonging to two or more cores. Furthermore, a constitutional formula may contain statements about free valence electrons at some of the cores.

In the computer-assisted solution of chemical problems and chemical documentation the chemical constitution of molecular systems is represented by connectivity lists or matrices.[1] Such computer-oriented representations of chemical constitutions are only unequivocal if they are based upon a canonical order of the atoms. In order to avoid ambiguities in the case of molecules which are representable by two or more resonance formulas, the same representation should result for a given molecule, regardless of the resonance structures considered.

A molecule may contain constitutionally equivalent atoms. This fact is reflected in NMR spectra, if the fine structure of the bands is neglected which is due to spin–spin coupling and stereochemical effects. An NMR spectrum then corresponds to the classes of constitutionally equivalent atoms whose nuclei are observed. As a rule, the relative areas of the peaks in the chemical shift pattern of ^1H NMR spectra correspond to the number of protons in the corresponding class of constitutionally equivalent atoms.

A molecule which contains constitutionally equivalent atoms is called constitutionally symmetric. Perception of constitutional symmetry is also important in the solution of chemical problems with the aid of computers. In computer-assisted synthetic design[2] neglect of constitutional symmetry leads to redundancies in the tree of synthetic pathways. Furthermore, a canonical order of the atoms in a molecule is readily generated if constitutional symmetry is adequately taken into account. A canonical order of the atoms in molecules is necessary for computer-oriented descriptions of molecules.

Several algorithms for generating unique representations of molecules have been reported so far,[3] but they were designed for the use in retrieval systems and do not directly produce any additional information which could possibly be used in systems for synthesis planning. The sole purpose of these algorithms is to generate somehow a unique description for a molecule. For another approach to this problem see ref 4. In this paper an algorithm is presented which detects the constitutional symmetry and generates a unique description of the molecule.

Constitutional Symmetry

Usually the symmetry of molecules is discussed in terms of point groups. Since the symmetry of three-dimensional objects is described, the point groups are based on the group of all orthogonal transformations in three-dimensional space, like rotations, reflections, and translations. A symmetry operation representing a certain symmetry property of a molecule sends

0002-7863/78/1500-0037$01.00/0

each atom onto the place of an atom of the same kind. Thus a symmetry operation describes the symmetry of a molecule with respect to three-dimensional space. No space, however, is necessary to describe the constitutional symmetry of a molecule. The constitutional symmetry may therefore be considered a more general symmetry than that described by point groups. In order to present the concepts of constitutional symmetry the constitutional formula of a compound is regarded as a graph. Atoms are represented by the vertices and the bonds are depicted by the edges of the graph. Furthermore, each vertex is labeled by the chemical element symbol of the corresponding atom. Free electrons and charges at the atoms may also be included in the labels of the vertices. The resulting finite, connected, undirected, and labeled graph[5] is henceforth called a molecular graph.

Let $\mathbf{V} = \{v\}$ be the set of vertices, and $\mathbf{L} = \{l\}$ be the set of labels assigned to the vertices in a molecular graph

$$M = [\mathbf{A},\mathbf{B}] \tag{1}$$

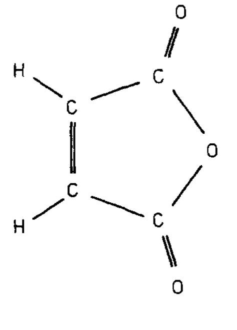
representing the constitution of a molecule.

$$\mathbf{A} = \{(v,l){:}\ v \in \mathbf{V}, l \in \mathbf{L}\} \tag{2}$$

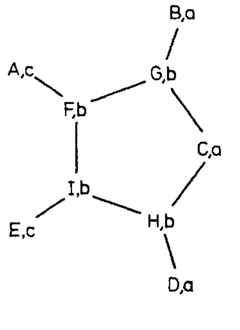
is a set of ordered pairs. The first coordinate is a vertex of the graph and the second one is the label assigned to it. Since the labels l identify chemical elements, a pair (v,l) denotes an atom in the corresponding molecule.

$$\mathbf{B} = \{(u,v){:}\ u \in \mathbf{V}, v \in \mathbf{V}\} \tag{3}$$

is a set of unordered pairs indicating which vertices are adjacent or which atoms are bonded in the molecule. As an example consider maleic acid anhydride. The structural formula I is represented by the molecular graph II.



I



II

The sets **V**, **B**, **L**, and **A** are **V** = {A, B, C, D, E, F, F, G, H, I}, **B** = {(A,F), (F,G), (F,I), (G,B), (G,C), (C,H), (H,D), (H,I), (E,I)}, **L** = {a, b, c}, **A** = {(A,c), (B,a), (C,a), (D,a), (E,c), (F,b), (G,b), (H,b), (I,b)}.

Let

$$\mathbf{S}(v,r,n) = \{(u, \mathbf{P}(u,n)){:}\ r = d(v,u)\} \tag{4}$$

be an ordered set of ordered pairs $(u,\mathbf{P}(u,n))$ formed by vertex u and its descriptor $\mathbf{P}(u,n)$. A descriptor $\mathbf{P}(v,n)$ assigned to vertex v is given by the ordered set

$$\mathbf{P}(v,n+1) = \{\mathbf{S}(v,r,n){:}\ r = 0,1,2\ldots\} \tag{5}$$

Equations 4 and 5 constitute a recursion relation for generating the descriptors $\mathbf{P}(v,n)$ for a vertex in the molecular graph. By assigning values to $\mathbf{P}(v,0)$ descriptors $\mathbf{P}(v,n)$ can be generated for increasing values of n. The initial values for the descriptors are taken to be the labels of the vertices.

The first coordinate of the ordered pairs $(u,\mathbf{P}(u,n))$ in eq 4 is a vertex u separated from vertex v by the graph theoretical distance $r = d(v,u)$. The graph theoretical distance is the length of the shortest path from vertex v to a vertex u, or just the smallest number of bonds which separate the corresponding atoms. For $r = 0$ u equals v and $\mathbf{S}(v,0,n)$ contains only $(v,\mathbf{P}(v,n))$. The vertices in the pairs $(u,\mathbf{P}(u,n))$ are adjacent to v if $r = 1$, e.g., S(G,1,0) contains (C,a), (B,a), and (F,b), if we consider vertex G in II. If r is greater than its maximal value for a certain vertex in a graph then $\mathbf{S}(v,r,n)$ is empty. The maximal value of r for, say vertex C, is three, whereas it is four for vertex B in II.

$\mathbf{S}(v,r,n)$ is defined as an ordered set. Let the order in $\mathbf{S}(v,r,n)$ be given by the second coordinates in the ordered pairs $(u,\mathbf{P}(u,n))$, i.e., by the order of the descriptors assigned to the vertices: $(u,\mathbf{P}(u,n)) < (v,\mathbf{P}(v,n))$, if $\mathbf{P}(u,n) < \mathbf{P}(v,n)$, and $(u,\mathbf{P}(u,n) = (v,\mathbf{P}(v,n))$, if $\mathbf{P}(u,n) = \mathbf{P}(v,n)$. E.g., in II the set of ordered pairs containing the first neighbors of vertex G and their initial descriptors is given by S(G,1,0) = {(C,a), (B,a), (F,b)}, if the order relation in the set of labels is taken to be the alphabetical order.

The descriptors $\mathbf{P}(v,n)$ are then constructed according to eq 5 by including the second neighbors, third neighbors, etc., of the vertex v. The descriptor for, say, vertex A in Il is thus given by P(A,1) = {{(A,c)}, {(F,b)}, {(G,b), (1,b)}, {(B,a), (C,a), (H,b), (E,c)}, {(D,a)}}. If notation is changed for brevity, the descriptors $\mathbf{P}(v,1)$ for the vertices in 11 are given by

P(A,1) ← c,b,bb,aabc,a
P(B,1) ← a,b,ab,bbc,ac
P(C,1) ← a,bb,aabb,cc
P(D,1) ← a,b,ab,bbc,ac
P(E,1) ← c,b,bb,aabc,a
P(F,1) ← b,bbc,aabc,a
P(G,1) ← b,aab,bbc,ac
P(H,1) ← b,aab,bbc,ac
P(1,1) ← b,bbc,aabc,a

In order to be able to generate the descriptors $\mathbf{P}(v,2)$, an order of the descriptors $\mathbf{P}(v,1)$ has to be defined, as it was necessary to assume an order of the initial descriptors or labels of the vertices. Let the order of the descriptors $\mathbf{P}(v,n)$ be given by their lexicographical order, i.e., the elements of two descriptors $\mathbf{P}(u,n)$ and $\mathbf{P}(v,n)$ are compared from left to right in much the same way as one proceeds if a word is looked up in a dictionary. Since the elements in a descriptor are sets $\mathbf{S}(v,r,n)$, an order of these sets must be defined first.

Let the sets $\mathbf{S}(v,r,n)$ be lexicographically ordered such that $\mathbf{S}(u,r,n) < \mathbf{S}(v,r,n)$, if some element in $\mathbf{S}(u,r,n) <$ the corresponding element in $\mathbf{S}(v,r,n)$, and all elements to the left in $\mathbf{S}(u,r,n)$ and $\mathbf{S}(v,r,n)$ are equal. E.g., S(C,2,1) = aabb precedes S(F,2,1) = aabc in 1I. The elements in the ordered sets $\mathbf{S}(u,r,n)$ and $\mathbf{S}(v,r,n)$ are compared until an order can be established, or the last element in the set with the smaller number of elements has been compared to the corresponding element in the set with the greater number of elements. If, say, $\mathbf{S}(u,r,n)$, contains more elements than $\mathbf{S}(v,r,n)$, and all elements in $\mathbf{S}(v,r,n)$ are equal to the corresponding elements in $\mathbf{S}(u,r,n)$, then $\mathbf{S}(u,r,n) < \mathbf{S}(v,r,n)$. E.g., S(G,2,1) = bbc precedes S(E,2,1) = bb in 11. Two sets $\mathbf{S}(v,r,n)$ are equal, if all corresponding elements are equal, and both sets have the same number of elements. Permutations of equal elements in the sets $\mathbf{S}(v,r,n)$ do not affect the order of the sets.

An order of the descriptors $\mathbf{P}(v,1)$ in 11 can now be established by first comparing the sets $\mathbf{S}(v,0,0)$. This leads to the order P(B,1), P(C,1), P(D,1) < P(F,1), P(G,1), P(H,1), P(1,1) < P(A,1), P(E,1). The comparison of the sets $\mathbf{S}(v,1,0)$, $\mathbf{S}(v,2,0)$, etc., finally leads to the following partial order of the descriptors $\mathbf{P}(v,1)$: P(C,1) < P(B,1), P(D,1) < P(G,1), P(H,1) < P(F,1), P(1,1) < P(A,1), P(E,1).

The descriptors $\mathbf{P}(v,1)$ are now used to construct the sets $\mathbf{S}(v,r,2)$ which in turn are taken to generate the descriptors $\mathbf{P}(v,2)$. The descriptors are thus recursively generated until at a level $n + 1$ the number of different descriptors is not greater than that at level n. If this condition is met, the descriptors $\mathbf{P}(v,n)$ last generated are called $\mathbf{P}(v,m)$. The descriptors $\mathbf{P}(v,m)$ are the crucial result of the algorithm since they reflect the constitutional symmetry of the atoms in a molecule. For maleic acid anhydride the algorithm stops after the first iteration because the descriptors $\mathbf{P}(v,2)$ do not further differentiate among the vertices. Since vertices A and E get equal descriptors the pairs (A,c) and (E,c) in II represent constitutionally equivalent atoms, as well as (F,b) and (I,b), (G,b) and (H,b), and (B,a) and (D,a), respectively. Constitutionally equivalent atoms are then given by the sets

$$\mathrm{E}(\mathbf{P}(v,m)) = \{e: e \in \mathbf{A}\} \qquad (6)$$

where each vertex in e has the same descriptor $\mathbf{P}(v,m)$.

The procedure for detecting constitutional symmetry can be concisely written in algorithmic form:

```
C10: Generate $\mathrm{S}(v,r,n)$, $r = 1,2,3 \ldots$
C20: Generate $\mathbf{P}(v,n + 1)$.
C30: Are there not more different $\mathbf{P}(v,n + 1)$ than $\mathbf{P}(v,n)$?
     If yes then stop.
C40: Set $n \leftarrow n + 1$.
C50: Go to C10.
```

In order to give some additional examples algorithm C has been used to determine the constitutional equivalence of the atoms in some tricyclodecanes. The results are shown in Figure 1.

It still must be demonstrated that vertices with equal descriptors $\mathbf{P}(v,m)$ represent constitutionally equivalent atoms. The chemical constitution of a molecule is given by the molecular graph. Its properties can be equated with the constitutional properties of the molecule.

Two atoms a and b are constitutionally equivalent, if there is a one-to-one correspondence $a \leftrightarrow b$ of atoms in the molecule such that the neighbors of each atom are preserved, or, more formally, if the vertices u and v, standing for a and b, are similar with respect to an automorphism[5] of the molecular graph.

The descriptors $\mathbf{P}(v,n)$ for all vertices v contain all the information pertaining to the constitution of a molecule. The first element in $\mathbf{P}(v,n)$ represents the atom v and all other elements represent the other atoms of the molecule ordered according to the number of bonds which separate them from the atom denoted by v and the kind of atom. The kind of atom is determined by the label l assigned to each vertex.

A descriptor $\mathbf{P}(v,n)$ is a representation of the molecular graph in which adjacency is preserved and adjacent vertices are put in the order of their descriptors $\mathbf{P}(v,n - 1)$. Comparing the descriptors $\mathbf{P}(v,m)$ is therefore nothing else than looking for a one-to-one correspondence of vertices by comparing their descriptors at the appropriate level of adjacency. If two descriptors $\mathbf{P}(u,m)$ and $\mathbf{P}(v,m)$ are equal, such a one-to-one correspondence of vertices has been found, and u and v are similar vertices in an automorphism of the molecular graph. This means, by definition, that the atoms associated with u and v are constitutionally equivalent.

Two atoms are not constitutionally equivalent, if there is at least one atom for which there is no related atom having the same neighbors under a one-to-one correspondence of atoms. The vertices representing these two atoms are then not similar with respect to an automorphism of the molecular graph because their descriptors $\mathbf{P}(v,m)$ differ in at least one level of adjacency. Therefore atoms being not constitutionally equivalent always get different descriptors.

Canonical Ordering

For the unequivocal representation of molecules by connectivity lists or related matrices[6] it is necessary to establish a canonical order of the atoms in a molecule using the information contained in the constitutional formula. The procedure for this purpose can only be based on invariant properties of the molecular graph. Especially the initial order of the atoms in a molecule cannot be employed in any way.

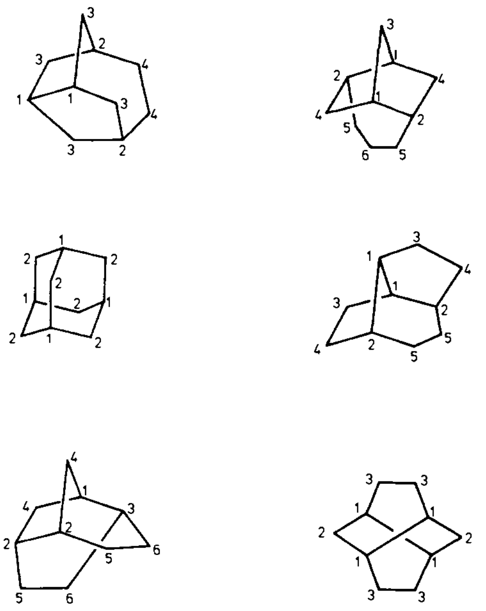


Figure 1. Constitutionally equivalent atoms of some tricyclodecanes.

Let k denote the number of times algorithm C has been invoked when generating a canonical order, and let

$$\mathrm{F}(k) = \{(v,\mathbf{P}(v,n))\} \qquad (7)$$

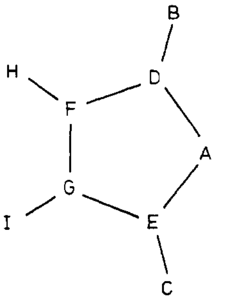
be an ordered set of ordered pairs not having unique second coordinates $\mathbf{P}(v,n)$ at iteration $n \geq m$. The order in $\mathrm{F}(k)$ is given by the second coordinates.

If $\mathrm{F}(k)$ is empty, the molecular graph is an identity graph[5] because there exists only one automorphism which is the identity map from $\boldsymbol{M}$ onto $\boldsymbol{M}$. In this case an unique order of the vertices has been found to be given by the uniqueness of their descriptors $\mathbf{P}(v,m)$. The atoms are then ordered according to the order of the descriptors $\mathbf{P}(v,m)$ of the corresponding vertices. Examples of molecules which can be represented by an identity graph are hydrogen cyanide and 2-chloropyridine.

If the set $\mathrm{F}(k)$ is not empty, the molecule contains constitutionally equivalent atoms, and it is not possible to totally order the set of vertices on grounds of algorithm C. The reason is that for the molecular graph then there exist more than one automorphism. In the case of II $\mathrm{F}(1) = \{(\mathrm{B},\mathbf{P}(\mathrm{B},1)), (\mathrm{D},\mathbf{P}(\mathrm{D},1)), (\mathrm{G},\mathbf{P}(\mathrm{G},1)), (\mathrm{H},\mathbf{P}(\mathrm{H},1)), (\mathrm{F},\mathbf{P}(\mathrm{F},1)), (\mathrm{I},\mathbf{P}(\mathrm{I},1)), (\mathrm{A},\mathbf{P}(\mathrm{A},1))$, and $(\mathrm{E},\mathbf{P}(\mathrm{E},1))\}$.

A total order in the set of vertices can be established by reinitiating algorithm C according to the following rule. A descriptor $\mathbf{P}'(v,n)$ is assigned to a vertex v contained in $\mathrm{F}(k)$ having a minimal descriptor $\mathbf{P}(v,n)$ in $\mathrm{F}(k)$. The descriptor $\mathbf{P}'(v,n)$ is chosen such that it is an immediate predecessor of $\mathbf{P}(v,n)$. The pairs $(\mathrm{B},\mathbf{P}(\mathrm{B},1))$ and $(\mathrm{D},\mathbf{P}(\mathrm{D},1))$ are minimal in $\mathrm{F}(1)$ of II. By assigning $\mathbf{P}'(\mathrm{B},1)$ to vertex B the equality of the descriptors for vertex B and D is removed since $\mathbf{P}'(\mathrm{B},1)$ now precedes $\mathbf{P}(\mathrm{D},1)$.

Algorithm C is then executed again. If the resulting descriptors $\mathbf{P}(v,n)$ are all different, the procedure stops. Otherwise algorithm C is applied repeatedly after assigning $\mathbf{P}'(v,n)$ to a vertex v which is minimal in $\mathrm{F}(k + 1)$. Algorithm C is reinitiated until all descriptors of the vertices are different. The vertices are then ordered according to the order of their last descriptors $\mathbf{P}(v,n)$. For generating a canonical order of vertices in II algorithm C must be reinitiated only once which leads to the order C < B < D < G < H < F < I < A < E of the vertices. F(2) is empty since no two vertices get equal descriptors at the next iteration. The vertices of the molecular graph II are thus renamed as shown in III.



III

Generating a total order in the set of vertices of a molecular graph is thus given by the following algorithm:

```
U10: Generate F(k).
U20: Is F(k) empty? If yes go to U70.
U30: Assign P'(v,n).
U40: Execute algorithm C.
U50: Set k ← k + 1.
U60: Go to U10.
U70: Put vertices in order of their descriptors.
```

If the molecule contains no constitutionally equivalent atoms, the uniqueness of the order of the atoms is given by the total order in the set of vertices. Since vertices get always different descriptors $\mathbf{P}(v,m)$ if they represent atoms which are not constitutionally equivalent, it must be proven that the descriptors are always assigned to the vertices in the same order, independent of the initial order of the vertices. The initial order does not enter the algorithms at any point. Therefore the order of the vertices is clearly independent of their initial order. For a given order relation $\leq$ in the set of labels $\mathbf{L}$ there exists only one descriptor $\mathbf{P}(v,m)$ which determines the position of the vertex v. The descriptors $\mathbf{P}(v,m)$ are not unique because at any level n vertices with equal descriptors can be permuted in $S(v,r,n-1)$ without affecting the position of the vertex v. Since there exists only one descriptor for each vertex at any level n determining the position of the vertex, the order of the vertices is unique when ordered according to their descriptors $\mathbf{P}(v,m)$.

Atoms in molecules having constitutionally equivalent atoms are ordered by reinitiating algorithm C. A descriptor $\mathbf{P}'(v,n)$ $n \geq m$ is assigned to a vertex v having a descriptor $\mathbf{P}(v,n)$ which is minimal in $\mathbf{F}(k)$. The order of descriptors already uniquely assigned to vertices is never affected by the descriptor $\mathbf{P}'(v,n)$ since it is never assigned to fixed vertices of an automorphism. Since there are at least two vertices with a minimal descriptor, the order of the vertices depends on which of the vertices with minimal descriptor is assigned the descriptor $\mathbf{P}'(v,n)$. Assigning a descriptor $\mathbf{P}'(u,n)$ to a vertex u establishes an order for all other vertices similar with respect to an automorphism. Another order results if a descriptor $\mathbf{P}'(v,n)$ is assigned to another vertex v. Because u and v are similar vertices the two orderings of the vertices induced by $\mathbf{P}'(u,n)$ and $\mathbf{P}'(v,n)$ differ by a permutation of vertices which is an automorphism of $\boldsymbol{M}$, i.e., an isomorphism of $\boldsymbol{M}$ onto itself. Therefore the two orderings represent the same graph and can be used interchangeably. Descriptors $\mathbf{P}'(v,n)$ are introduced until all symmetry has been destroyed and a total order of the vertices results. A number of canonical orderings can be generated in the described manner, but all of them represent the same molecular graph. Therefore any such ordering is equally well suited to uniquely describe the molecule represented by the molecular graph.

Unique Molecular Descriptors

Let the vertices of a molecular graph be canonically ordered. A unique representation is then readily constructed for the corresponding molecule. The representation must, of course, comprise the information contained in the constitutional formula. Let

$$\boldsymbol{D} = [\mathbf{A},\mathbf{B}] \qquad (8)$$

be the representation of a molecule, where $\mathbf{A}$ and $\mathbf{B}$ have the same meaning as in eq 1. This representation is called the molecular descriptor. As long as only molecules having no stereoisomers are considered, $\boldsymbol{D}$ is unique.

So far the discussion has been limited to the constitution of molecules, and consequently $\boldsymbol{D}$ cannot be used to distinguish stereoisomers of a compound. Since the configurational aspects of molecules can be treated independently from a specific order of the atoms in a molecule, it is possible to merely extend the molecular descriptor by a descriptor for the stereochemical features of a molecule. As has been shown by Ugi[6] et al., the stereochemical features mostly needed in synthesis planning can be represented by parities of chiral centers with three and four ligands. The molecular descriptor

$$\boldsymbol{Z} = [\mathbf{G},\mathbf{B}] \qquad (9)$$

is different for stereoisomers.

$$\mathbf{G} = \{(v,l,p) : v \in \mathbf{V}, l \in \mathbf{L}, p \in \{-1, 0, +1\}\} \qquad (10)$$

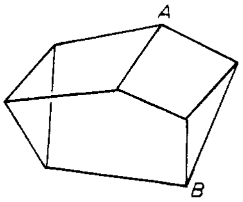
is a set of ordered triplets where v is a vertex, l its label identifying the chemical element, and p indicating the parity of the atom represented by vertex v and label l. The parity p can assume the values -1 and $+1$ for chiral centers depending on the relative configuration of the ligands, and it is 0, if the atom is no chiral center.

Discussion

$\boldsymbol{D}$ and $\boldsymbol{Z}$ have a generic structure. Not every part has to be checked, if two different descriptors $\boldsymbol{Z}$ and $\boldsymbol{Z}'$ are to be compared. If $\mathbf{G}$ and $\mathbf{G}'$ are not equivalent, the molecules are different and no other component of $\boldsymbol{Z}$ is needed for further comparison. For $\boldsymbol{Z}$ and $\boldsymbol{Z}'$ differing only in certain elements of $\mathbf{G}$ and $\mathbf{G}'$, the molecules are still different, but have similar constitutional formulas, e.g., esters and thioesters. Specifying $\mathbf{G}$ with all parities being zero identifies all structural isomers of a molecular formula, like benzene, Dewar benzene, and prismane for $(CH)_6$. Molecular descriptors differing only in the parities of the atoms denote stereoisomers.

The canonical order of vertices in a molecular graph and therefore the molecular descriptor still depend on the initial labels and the order relation $\leq$ actually used. Usually the chemical symbols of the atoms are written at the vertices of a structural formula and could be taken as labels. The order relation $\leq$ in the set of labels then could denote alphabetical order. Since the algorithms are designed for use in a computer, it is more convenient to replace the symbols by the atomic numbers and take the order of natural numbers as the order relation. Sometimes it may be necessary to distinguish between isotopically labeled compounds. Different molecular descriptors can be generated for them by considering the atomic weights in addition to the atomic numbers when assigning initial labels. If no information pertaining the valence electrons of the atoms enters the initial labels, molecules having the same molecular graph but differing in the number or distribution of electrons, or both, get the same descriptor. This is especially true for resonance structures in which the electrons may be formally distributed in several ways. Generating the same molecular descriptor for resonance structures may be quite advantageous for some applications as documentation. Regarding all resonance structures of a compound as different entities is highly impracticable, if one takes into account the large number of resonance structures which can be.drawn for even fairly small molecules. On the other hand, the algorithms described are designed in a way that electronic information like the number of valence electrons of each atom can be handled by including it in the initial labels. This may be necessary, if molecular descriptors for charged species are to be generated. There is in fact no restriction on what kind of initial labels is used, as long as the same kind of atom is associated with the same label, and the order relation ≤ is the same for all molecules for which a name is generated.

The reported algorithm works for all chemical structures usually encountered in chemical synthesis. Because of the nature of the atomic descriptors defined there are some exceptional structures in which the maximal numbers of sets of constitutionally equivalent atoms are not found. In the following example atoms A and B are recognized as constitutionally equivalent. They are, however, not equivalent. If such structures are encountered some additional conditions, like the number of ring closures at a certain distance from a vertex, have to be considered to distinguish the vertices. In order to keep the algorithm as fast as possible this feature has not been included.



IV

The algorithms have been implemented in FORTRAN IV and Pl/1 and have been extensively used in our synthesis planning program. The fact that constitutional symmetry can be detected leads to a substantial reduction of the number of precursors in the reaction network because generating duplicate precursors can be avoided. The number of precursors is further reduced by the fact that mesomeric structures always get the same descriptor. No precursor is regarded as a different compound merely because the electrons are formally distributed in some other way.

Acknowledgment. We gratefully acknowledge the support of this work by the EC.

References and Notes

(1) (a) H. J. Hiz, *J. Chem. Doc.*, **4**, 173–180 (1964); (b) L. Spialter, *J. Am. Chem. Soc.*, **85**, 2012–2013 (1963); (c) L. Spialter, *J. Chem. Doc.*, **4**, 261-268 (1964); (d) *ibid.*, **4**, 269-273 (1964); (e) E. Meyer, *Angew. Chem.*, **82**, 605–611 (1970).

(2) (a) I. Ugi, J. Gasteiger, J. Brandt, J. F. Brunnert, and W. Schubert, *IBM Nachr.*, **24**, 185–189 (1974); (b) I. Ugi, *ibid.*, **24**, 180–184 (1974); (c) J. Brandt, J. Friedrich, J. Gasteiger, C. Jochum, W. Schubert, and I. Ugi, Proceedings of the American Chemical Society Centennial Meeting, New York, N.Y., 1976; (d) J. Brandt, J. Friedrich, J. Gasteiger, C. Jochum, W. Schubert, and I. Ugi, Proceedings of the III International Conference on Computers in Chemical Research, Education, and Technology, Caracas, Venezuela, 1976.

(3) (a) H. L. Morgan, *J. Chem. Doc.*, **5**, 107–113 (1965); (b) R. H. Penny, *ibid.*, **5**, 113–117 (1965); (c) M. Randic, *J. Chem. Inf. Comput. Sci.*, **15**, 105–108 (1975); (d) W. T. Wipke and T. M. Dyott, *J. Am. Chem. Soc.*, **96**, 4825–4833 (1974).

(4) C. Jochum and J. Gasteiger, *J. Chem. Inf. Comput. Sci.*, **17**, 113 (1977).

(5) (a) F. Harary, "Graph Theory", Addison-Wesley, Reading, Mass., 1972, pp 161–163; (b) W. Wagner, "Graphentheorie", B. I. Wissenschaftsverlag, Mannheim, Germany, 1970, pp 100–105.

(6) J. Blair, J. Gasteiger, C. Gillespie, P. D. Gillespie, and I. Ugi, *Tetrahedron*, **30**, 1845–1859 (1974).

Optimized Geometries of the Saddle-Point Rotamers of Formamide

Roman F. Nalewajski[1]

Contribution from the Department of Chemistry, University of North Carolina, Chapel Hill, North Carolina 27514. Received March 22, 1977

Abstract: The ab initio 4-31G fully optimized geometries of the saddle-point rotamers of formamide are reported, with the NH_2 group twisted around the CN bond by 90 and 270°, respectively. The effects of geometry relaxation are discussed in terms of geometry distortions as well as changes in the energy components, dipole moment, and the net charges on atoms. The most important relaxational coordinates were found to be the CN bond length (increase by about 0.05 Å) and the HNH angle (increase by about 13 and 18°, respectively). The predicted variations of these geometrical parameters appear to be consistent with the main change in the electronic structure accompanying the rotation, namely, a lack of delocalization of the nitrogen lone pair into the CO π system in both orthogonal conformers. The energy drops obtained due to geometry optimization (5.64 and 6.24 kcal/mol, respectively) suggest that the rigid-rotation assumption was much responsible for overestimation of the previously reported 4-31G rotational barrier in formamide. Comparisons are made between the ab initio and MINDO geometry relaxational effects, in order to check the validity of the semiempirical SCF MO predictions. The MINDO method was found to fail to predict correctly the lone pair effects in bonded intereactions.

Introduction

The investigation of the geometry distortions accompanying inversion and internal rotation in molecular systems is of growing significance in quantum chemistry.[2] Geometry optimization is essential for reliable assignment of the equilibrium conformers[3-5] and has a considerable effect on the predicted barriers and their components, e.g., ref 4 and 5. In recent years a number of efficient gradient-type procedures have been developed and applied both within ab initio and semiempirical methods.[8,9] Energy gradients are directly provided by the ZDO-type semiempirical methods. In ab initio calculations they must be obtained by a finite difference technique, but the effort is also being made to develop methods for direct calculation of energy gradients.[10] The literature[2,11] already contains many successful applications of the ab initio calculations for