

## Distances in Molecular Graphs<sup>†</sup>

Wataru Katouda, Takashi Kawai, Tetsuhiko Takabatake, and Akio Tanaka

Organic Synthesis Research Laboratory, Sumitomo Chemical Company 1-98, Kasugade-naka 3-chome, Konohana-ku, Osaka 554-8558, Japan

Malcolm Bersohn\* and Daniel Gruner

Department of Chemistry, University of Toronto, Toronto, Ontario M5S 3H6, Canada

Received: February 16, 2004; In Final Form: June 15, 2004

This paper discusses the finding of vertex to vertex distances in molecular graphs. Having found these distances, one can obtain a method for canonical numbering of the atoms in a molecule, which depends on the atomic properties and the distances between equivalence classes. This does not use the traditional Morgan algorithm. Using distances one can also perceive rings. Finally, substructures of interest can be detected using distances between the central atoms of various functional groups. The set of vertex distances are thus a kind of lens for examination of the graph properties of molecules. Applications have thus far been only in organic chemistry. Application to physical chemistry may appear wherever molecular graphs can be helpful, such as in calculations concerning molecules of high symmetry.

### 1. Introduction

Ring perception in molecules, and graph classification of the atoms therein, is of obvious importance in organic chemistry applications. It is also useful in molecular modeling, where the graph structure of molecules defines the force fields employed for the molecular mechanics calculations underlying the models. This paper describes an algorithm for finding vertex to vertex distances in molecular graphs. This is then followed by the canonical numbering of the atoms, which depends on the atomic properties and the distances between equivalence classes.

The mathematical concept called a graph has only edges and vertexes. Each edge connects two vertexes. There is no concept of length or angle. We can describe any graph by means of a connection table. In a connection table, after numbering the vertexes of a graph, we can for each vertex list the numbers of the vertexes to which it is connected to by edges. In the graph representation of a molecule, a chemical bond is taken as an edge, the atoms as vertexes. Equivalence classes of atoms under graph automorphisms can be found, as well as rings. The shortest path between atom  $x$  and atom  $y$  in the molecule is the path with the fewest edges. The graph distance between any two atoms  $x$  and  $y$  is the number of edges on the shortest path connecting  $x$  and  $y$ . Conventionally, hydrogen atoms are not mapped onto vertexes but rather are viewed as local properties of the atom to which they are bonded. Also we assume that we are dealing with a connected graph, i.e., a graph in which every vertex is connected to every other vertex via a series of edges. Herein we present new methods for canonically ordering the atoms of a molecule and for perceiving the rings in a molecule, both depending on the graph distance properties in the molecule.

### 2. Methods for Finding Atom-to-Atom Distances in a Connected Graph

There are two such methods. Both of them find the distances in ascending order of size. The atoms can be numbered

arbitrarily. The distance values of unity are already displayed in the adjacency matrix, which lists the numbers of neighbors of each atom.

We wish to find  $d(x,y)$ , the distance in edges between atoms  $x$  and  $y$ . The well-known method is the matrix multiplication method. It is based on the powers of the adjacency matrix. The latter is a symmetric square matrix,  $\mathbf{A}$ , in which  $A_{x,y} = A_{y,x} = 1$  if  $x$  and  $y$  are neighbors in the molecule. Otherwise  $A_{x,y} = A_{y,x} = 0$ .  $\mathbf{A}$  is of size  $n \times n$ , where  $n$  is the number of atoms in the molecule. Consider the equation

$$\mathbf{A} \times \mathbf{A} = \mathbf{A}^2$$

If  $A_{x,j} = 1$  and  $A_{x,y} = 0$  and  $A_{j,y} = 1$ , then  $A_{x,y}^2 = 1$ .

This is clear because  $x-j-y$  form a string. Because  $x$  is not adjacent to  $y$ ,  $x-j-y$  cannot form a three-membered ring. If we consider what is involved in the matrix multiplication of  $\mathbf{A}$  by  $\mathbf{A}$  then we know that the sum

$$\sum_{k=1}^n A_{x,k} A_{k,y}$$

must be nonzero. This is so because there is at least one term in the sum, i.e.,  $A_{x,j} A_{j,k}$ , which is nonzero. We conclude that the distance from  $x$  to  $y$  is 2. We can state more generally the theorem that

For any power  $p$  greater than 1, of the matrix  $\mathbf{A}$ , if  $A_{x,y}^p$  is nonzero for this value of  $p$  but it was zero in  $\mathbf{A}^z$  for all values of  $z$  from 1 to  $p - 1$ , then the distance from  $x$  to  $y$  is  $p$ .

Proof: The elements of  $\mathbf{A}^p$  are

$$\sum_{k=1}^n A_{x,k}^p A_{k,y}$$

Let us suppose the following:

$A_{x,y}^z$  is zero for all values of  $z$  from 1 to  $p - 1$ .

There exists an atom  $j$  that is a distance of unity from  $y$ . The theorem holds for the distance  $p - 1$ .

<sup>†</sup> Part of the special issue "Richard Bersohn Memorial Issue".

\* Corresponding author. E-mail: mbersohn@chem.utoronto.ca.

In other words  $A_{xj}^z$  is zero for all values of  $z$  from 1 to  $p - 2$  but it is nonzero for  $z = p - 1$ .

Consider the product  $A_{xj}^{p-1}A_{jy}$ . This is

$$\sum_{j=1}^n A_{xj}^{p-1}A_{jy}$$

which is  $A_{xy}^p$ . Because  $A_{jy}$  is 1 and  $A_{xj}$  is nonzero then the value of  $A_{xy}^p$  must be nonzero. In this way we have a proof of the theorem by induction. The theorem is true for  $p = 2$ ; hence, it is proven. Unfortunately, this elegant algorithm requires many multiplications hence it is slow.

In the more rapid method that we devised, we examine all pairs of neighbors  $\{x, y\}$  such that  $x < y$ . All neighbors  $j$  of  $y$ , for which the distance  $d(j, x)$  is unknown must be separated from  $x$  by a distance of 2.  $d[x][j]$  must equal  $d[j][x]$  in an undirected graph; so they are both given the value 2. Similarly, all neighbors  $k$  of  $x$  for which the distance  $d(k, y)$  is unknown, must have corresponding distances  $d(k, y)$  and  $d(y, k)$  equal to 2. We next examine all pairs of atoms  $\{x, y\}$  such that  $x < y$  and  $x$  and  $y$  are separated by a distance of 2. The neighbors of  $x$  and  $y$  will similarly give us all distances of 3 that exist in the molecule. We proceed in this way until all the  $n(n - 1)/2$  distances are known. The underlying principle is that if  $d(x, y) = z$ , then any neighbor  $j$  of  $x$  the distance of which from  $y$  is hitherto unknown must have the value  $d(x, j) = z + 1$ . This is analogously true for any neighbor  $k$  of  $y$ . Inevitably, we find the distances in ascending order.

### 3. Distance Method for Numbering the Atoms of a Molecule Canonically

Strictly speaking, a system for storing or manipulating molecular structure does not need a uniform system for numbering the atoms of a molecule. In principle, one can always do an atom by atom comparison<sup>1,2</sup> so that benzene with atoms numbered from 1 to 6 in order around the ring can be mapped onto benzene in which the neighbors of atom 1 are 2 and 3 and the neighbors of 6, para to 1, are 4 and 5 (see Figure 1).

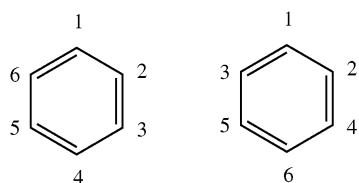


Figure 1.

However, there is a saving of time if we have a representation that is uniform by means of a set of rules. We can then more quickly answer the question “Are these two isomers the same molecule?” This is important for files such as those of Chemical Abstracts Service, which contain computer representations of the structures of more than 24,000,000 molecules. We can locate a set of isomers using an index based on the exact molecular weight. But a set of isomers may contain many molecules, the rapid distinguishing of which presents a problem to be solved. In our case, when our synthesis design program is searching for efficient routes, we frequently generate the structure representations of more than 2,000,000 molecules. It is a distinct advantage to be able rapidly to decide whether a structure previously generated with the same molecular formula is or is not the same molecule as the current isomer.

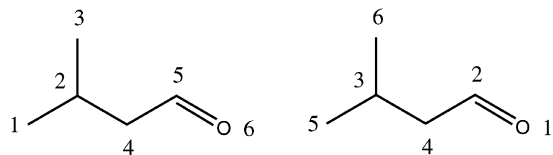
In our method we first divide the atoms into equivalence classes and next we decide the numbering of atoms within the

classes. The relationship between the current order of the atoms and their original numbering is preserved separately. The Morgan algorithm<sup>3</sup> is not used. In the Morgan algorithm there is a sum over properties of neighbors that is passed along step by step. This means that at first two atoms with the same atomic properties are differentiated if possible by the atomic properties of their neighbors. If the neighbors of atom  $x$  have the same atomic properties as the neighbors of  $y$ , then we reach further out to comparing the next nearest neighbors and so on until either we reach a differentiating environment at some distance  $d$  from the atoms  $x$  and  $y$  or we have explored the most distant atoms and conclude that the atoms are equivalent. We can then number the atoms according to a number derived from this search. For example the first  $\text{CH}_2$  carbon in  $\text{CH}_3\text{CH}_2\text{CHFCH}_2\text{-CH}_2\text{CH}_3$  is distinguished from all the other ones as it receives information from the nearby fluorine atom sooner than the third  $\text{CH}_2$  carbon. Similarly, it receives information from a methyl carbon sooner than does the second  $\text{CH}_2$  carbon. The Morgan algorithm has been criticized by Cieplak and Wisniewski<sup>4</sup> as being on occasion ambiguous so the same molecule can acquire two or more registration numbers.

Our assignment of the atoms to equivalence classes begins by dividing the atoms into various sets depending on their atomic property values. These sets are not necessarily the ultimate equivalence classes. We use an ad hoc “atomic property” based on (1) atomic number, (2) unsaturation value, (3) the number of attached hydrogen atoms, and (4) a functional group number. We first sort and rank the atoms in descending order of their atomic property values. The property value of an atom is the sum of the following four quantities: (1)  $10,000,000 \times$  atomic number, (2)  $100000 \times$  the unsaturation value, (3)  $10000 \times$  (4 - the number of attached hydrogen atoms), and (4) the functional group number.

The unsaturation value is 0 for a saturated atom, 1 for an atom in a six-membered aromatic ring, 2 for an atom in a carbon-carbon double bond, 4 for an atom doubly bonded to a heteroatom, 6 for an atom doubly bonded to two other atoms, and 8 for an atom triply bonded to another atom. The reason for using 4 minus the number of attached hydrogen atoms rather than the number of attached hydrogen atoms directly is our wish to give more substituted atoms a higher value. The atom with the highest value of the atomic property is numbered 1 and so on. The functional group number is not a strictly local property. An aldehyde carbonyl carbon is given the arbitrary value of 112, an ester carbonyl carbon gets the functional group value of 144, a ketone carbonyl carbon gets the functional group number 136. These numbers differ even though the local properties of the atoms are all the same. So the functional group number depends on what neighbors are present. In most functional groups only one atom gets the functional group number. The other atoms in the functional group are usually given the functional group number 0. For example, in triethylamine, only the nitrogen has a nonzero functional group number, i.e., 100. The other six non-hydrogen atoms all have the functional group 0. An exception to this scheme is the pairing of atoms with identical local properties, e.g., the oxygen atoms in a peroxide, the carbons in a  $\text{RCH}_x = \text{CH}_x\text{R}$  double bond where  $x$  is the same for both doubly bonded carbons, the doubly bonded nitrogens in an azo compound  $\text{RN}=\text{NR}'$ , etc. These functional group numbers, i.e., 100, 112, 136, and 144 are all arbitrary and are only a device to save time when sorting the atoms. Because they are arbitrary, they are of little general importance for other canonicalization algorithms and it is unnecessary to describe their assignment scheme in detail.

Let us take 3-methylbutanal as a specific example. Figure 2 shows the initial and final, canonical numbering. At the end of the sorting, the oxygen atom, originally numbered 6, will have successively exchanged values with atoms numbered 5, 4, 3, 2, and 1. Similarly, the atom originally numbered 1 will have exchanged values with atoms originally numbered 6, 5, 4, and 2. Atom number 3 will have successively exchanged values with atoms originally numbered 6, 5, 4, and 2.



**Figure 2.** 3-Methylbutanal, with initial numbering on the left and final, canonical, numbering on the right.

In this example, CH<sub>3</sub>, CH<sub>2</sub>, saturated CH, unsaturated CH, and unsaturated oxygen constitute five sets. Four of these sets have a unique atom and the methyl set has two equivalent members. In handling a more general molecule, we have to go further because there will be nonequivalent methyl groups, e.g., in cholesterol, and similarly more than one situation for saturated CH<sub>2</sub>, etc. In other words, the sets may consist of multiple equivalence classes.

The next task the algorithm prescribes at this point is the delineation of the boundaries of the sets. In the case of 3-methylbutanal we have the following set boundaries:

set no.	beginning	end		
1	1	1	O	unsaturated
2	2	2	CH	doubly bonded to a heteroatom
3	3	3	CH	saturated
4	4	4	CH <sub>2</sub>	saturated
5	5	6	CH <sub>3</sub>	methyl

When sets consist of a single atom, the atom is unique. The set is now recognized as an equivalence class. When sets consist of more than one atom, we have to find out if all the atoms in the set are equivalent, i.e., if it is an equivalence class or if the set has nonequivalent members.

The basic technique used here is to compare the distances of the atoms of multiatom sets to the atoms of other sets. In the case of 3-methylbutanal above, we find that the two methyl carbon atoms are equidistant from the saturated CH, equidistant from the methylene, in fact, equidistant from all the other atoms. Hence, the methyl groups are judged to be equivalent. To define it more generally, atoms  $x$  and  $y$  are equivalent if and only if for all sets  $Z$ , the set of the distances of  $x$  to the members of the set  $Z$  is the same as the set of the distances of  $y$  to the members of the set  $Z$ .

If  $x$  and  $y$  are not equivalent by the above measure then the atom numbered  $y$  must be removed from the set in which  $x$  is found. The atom  $y$  is now assigned to a new set.

At this point, the atoms are assigned to various atom sets. Inside of each set, all the atoms must be equidistant from the other sets. In other words, if  $x$  and  $y$  are atoms in the same atom set then the set of distances of atom  $x$  from any atom set  $Z$  must be the same as the set of distances of atom  $y$  from the same atom set  $Z$ . The precedence of the sets depends on their distances from the other sets. Initially, one atom set will have the highest ranking. If atoms in the same set are found to be nonequivalent, then they are placed in different sets with new boundaries for the old set and the new set. The process continues until we cannot find any more nonequivalent atoms in any set.

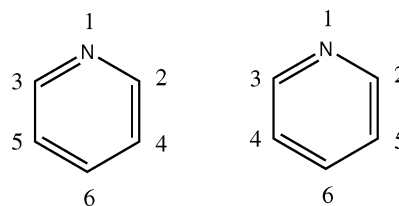
We now conclude that the sets are equivalence classes. For example, naphthalene initially looks like just two sets, the two ring junction atoms and the other eight atoms. But the latter have two different sets of distances from the ring junction atoms, thus dividing them into the classes of  $\alpha$  atoms and  $\beta$  atoms.

Graph distance has no relationship to geometry, hence none to stereochemistry. Here we need to introduce some nongraph distinctions for so far considered equivalent atoms. There are two-dimensional distinctions. Atoms associated with a double bond around which rotation is not possible, become nonequivalent if one atom is associated with a double bond with the *E* arrangement and the other atom is associated with a different double bond of the *Z* arrangement. The *E* arrangement atom is given precedence. Similarly, in three dimensions, chiral centers that are exchanged by a symmetry operation constitute an equivalent clockwise and counterclockwise pair, and the atom with clockwise arrangements of its ligands takes precedence. Also in a molecule such as 1-chloro-2,2-dimethylcyclohexane the methyl carbon that is on the same side of the ring as the chlorine atom gets a higher priority than the other methyl carbon. They are not in the same class.

Besides collecting the atoms into equivalence classes, we must also sort the connection table so that the classes appear in the connection table in monotonically descending order of precedence. The sort is based on a comparison. If two classes have different atomic properties, then they must be exchanged if the one with higher values of the atomic properties happens to be closer to the end of the connection table. If the two classes have equal values of the atomic properties, then the precedence is based on the sorted set of distances to the other classes. The class with smaller value of this sorted set of distances must have higher precedence. If the class with higher precedence happens to be lower down in the connection table, then the positions of the two classes in the connection table must be exchanged. This comparison and exchange process continues until the connection table contains the classes exactly in the descending order of precedence.

We believe that ours is the first working use of such a non-Morgan algorithm in a large scale chemical information system. We have not previously published this algorithm for canonically numbering the atoms of a molecule. The synthesis design program developed by the authors is in use by more than 100 users at the Sumitomo Chemical Corp.<sup>5</sup>

**3.1. Numbering the Neighbors.** Simply sorting the equivalence classes is necessary but not always sufficient to provide the canonical numbering. Consider for example a pyridine molecule. We have a choice of two numberings, as shown in Figure 3:



**Figure 3.**

We choose the lexicographically minimal numbering, which is that shown in the left structure above. This is easy enough when there are only two choices. But in anthracene there are 3456 choices. To generate them all and then find the lexicographically minimal one is more time-consuming than using the following rules.

**3.2. Connections between Atoms.** We wish to get the lexicographically minimal representation of the molecule. For example, in butane there is a two-membered set of CH<sub>2</sub> carbons and a two-membered set of CH<sub>3</sub> carbons. We must have atom number 1 connected to atoms numbered 2 and 3. Similarly, atom number 2 must be connected to atoms 1 and 4. The adjacency matrix, or connection table, is shown in Figure 4. The alternative, in Figure 5, is lexicographically larger than the first one.

Atom	Nbr 1	Nbr 2
1	2	3
2	1	4
3	1	0
4	2	0

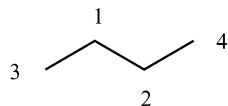


Figure 4.

Atom	Nbr 1	Nbr 2
1	2	4
2	1	3
3	2	0
4	1	0

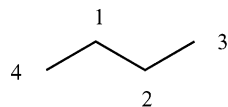


Figure 5.

How then should we obtain the lexicographically minimal numbering in a far more complex situation? First we must find out if there are rings in the molecule. The standard equation<sup>6</sup> is

$$\text{number of rings} = \text{number of edges} - \text{number of nodes} + 1$$

If the number of rings is zero, we have only to choose the minimum unused number. For example, see the case of triethylamine in Figure 6.

Atom	Nbr 1	Nbr 2	Nbr 3	
1	2	3	4	Nitrogen
2	1	5		CH <sub>2</sub>
3	1	6		CH <sub>2</sub>
4	1	7		CH <sub>2</sub>
5	2			CH <sub>3</sub>
6	3			CH <sub>3</sub>
7	4			CH <sub>3</sub>

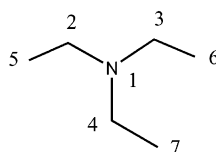


Figure 6.

When we are choosing the neighbors of atom 2, we see that the bond to atom 1 is already decided for it. It remains to choose an atom of the set {5, 6, 7}. The smallest unused atom is 5 so we use it. Similarly, atom 3 has already been assigned 1 as a neighbor, and the smallest hitherto unused neighbor of the set {5, 6, 7} is 6.

The case of a cyclic molecule is not so trivial. In this case we first examine the rings using the noncanonicalized connection table. The general rule is to select for the atom in an equivalence class the smallest unused compatible number. The content of the term compatibility will be explained later. Let us consider cyclohexane. All the atoms are in the same equivalence class {1, 2, 3, 4, 5, 6}. Clearly, pursuing our smallest unused number principle atom 1 must be bonded to atoms 2 and 3. This gives us the following partially complete connection table (Figure 7).

Atom	Nbr 1	Nbr 2
1	2	3
2	1	
3	1	
4		
5		
6		

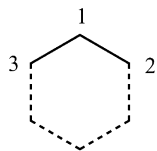


Figure 7.

Now we decide on the neighbors of atom 2. Atom 1 has already been chosen. We have a proviso that the smallest unused number in the appropriate equivalence class must be used if by so doing we do not form a ring that the original molecule lacked. In detail, if a certain numbering implies a ring, then the ring must be the same size as a ring known to be present in the molecule and contain the same number of atoms of the same equivalence classes arranged in the same order around the ring. This is what we mean by the term compatible. Otherwise, we must reject the numbering. For example, if we chose 3 as the neighbor of 2, then we would get Figure 8.

Atom	Nbr 1	Nbr 2
1	2	3
2	1	3
3	1	2
4		
5		
6		

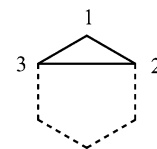


Figure 8.

This implies a three-membered ring that we know not to be present from a previous examination of the rings in the non-canonicalized connection table. Hence, instead of the smallest unused number, we use the smallest unused compatible number. In this case it is 4. With 4 as the neighbor of 2 we get Figure 9.

Atom	Nbr 1	Nbr 2
1	2	3
2	1	4
3	1	
4	2	
5		
6		

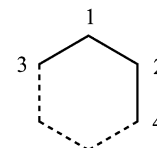


Figure 9.

Now we consider the other neighbor of atom 3 besides atom 1. It cannot be atom 2 because atom 2 has both of its neighbors assigned. So the smallest unused number is 4. However, 4 is not compatible as the resulting connection table implies a four-membered ring, 1-2-4-3 (Figure 10), and such a ring does

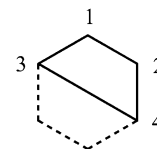


Figure 10.

not exist in the molecule, by our previous examination. Using the next smallest unused number, i.e., 5, we get Figure 11.

Atom	Nbr 1	Nbr 2
1	2	3
2	1	4
3	1	5
4	2	
5	3	
6		

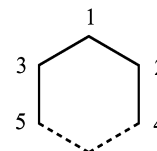


Figure 11.

As above, atom 5 would not suit as a neighbor for atom 4 because that would imply a nonexistent five-membered ring. We are forced to use the smallest remaining atom, the sole remaining atom, 6, as the second neighbor of atom 4. Finally, atom 5 must have a second neighbor and the only unused atom

is 6, obtaining thus the canonicalized connection table as shown in Figure 12.

Atom	Nbr 1	Nbr 2
1	2	3
2	1	4
3	1	5
4	2	6
5	3	6
6	4	5

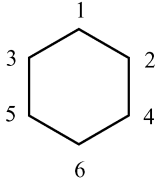


Figure 12.

#### 4. Simple Ring Perception Algorithm Using Atom–Atom Graph Distances

We present a simple algorithm for finding the smallest set of smallest rings of a molecular graph. Compared with conventional methods, the algorithm reduces blind traversal of the graph. It has an extra cost of a different kind, i.e., the time spent in calculating the graph distance between each pair of atoms in the molecule.

A monumental discussion of the problem of recognition of rings in computer representations of molecular structure and a summary of the literature on this subject prior to 1990 is in the papers of Lynch et al.<sup>7–9</sup> More recent advances are summarized by Figueras.<sup>10</sup>

The algorithm to be described here finds the smallest set of rings in a molecule such that in each ring of the set no proper subset of the ring atoms constitute a ring. The earliest algorithms for accomplishing the task of finding the smallest set of smallest rings of a molecular structure representation may be described as traversal algorithms. In one such algorithm, one travels on paths branching out from a particular atom until one of the paths meets the starting point again. Thus a ring is found. This might be described as a wandering algorithm.<sup>2,3,4</sup> The algorithm of Balducci and Pearlman<sup>11</sup> is distinguished from the previous ones in that the traversal is concurrent. At each atom in the graph one collects all possible paths from other atoms. When atom Y has received a path message from atom X and atom X has received a different path message from atom Y, then a ring is found. In the Balducci–Pearlman algorithm there are as many starting points as there are atoms in the molecule. Because the “broadcast messages”, i.e., paths, have monotonically increasing length, the smallest rings are found first. The work of Balducci and Pearlman<sup>11</sup> also established the exact complexity of their algorithm, a unique feat in this area.

Later Figueras<sup>10</sup> introduced the newest method for finding the smallest set of smallest rings. It is the most efficient algorithm devised so far. Like the Balducci–Pearlman it collects information at each atom simultaneously from all atoms in the molecule.

The algorithm we give here might be described as a “look ahead” algorithm. We do not traverse until we are sure we have discovered opposite atoms in a ring. In general, these opposite atoms are found before establishing the identities of any intervening atoms in the ring.

**4.1. Outline of the Method for Finding Odd-Membered Cycles.** The odd-membered ring will consist of  $2D + 1$  members, where  $D$  is some positive definite integer. For any pair  $(x, y)$  of neighboring atoms, the requirement that they are in a ring of size  $2D + 1$  implies that there must exist some atom  $z$  that is equidistant from  $x$  and  $y$ . Furthermore, the distance from  $z$  to  $x$  (and to  $y$ ) must be exactly  $D$ . The algorithm must guarantee that the ring shared by atoms  $x, y$ , and  $z$  is one of the smallest set of smallest rings. When  $D$  is 1, the search over all pairs of neighboring atoms guarantees that we have found all

of the three-membered rings. Next we find, by an analogous process for even-membered rings to be described below, all of the four-membered rings. After that we seek five-membered rings, etc. Throughout, we use the principle that we will not reuse any pair of atoms  $(x, y)$  if the bond between them is already present in a perceived ring.

For example, in Figure 13 the edge  $x - y$  has been used to find ring II; therefore it cannot be used to find ring III. To find ring III, we must start with a bond that has not yet been found to be cyclic, e.g., the edge  $y - b$ . Hence we will never find the compound ring consisting of the fusion of I and II. Ring III must be found starting from another pair, e.g.,  $y$  and  $b$ . Whenever there is no embedded ring in the molecule, each ring of the smallest set of smallest rings must contain at least one edge that is unique to it.

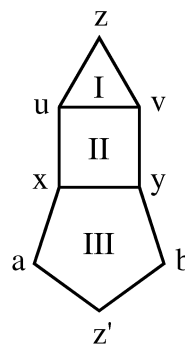


Figure 13.

Let us look at the task of finding the odd-membered rings such that  $D$  is greater than 1. We scan through the atoms of the molecule looking at all pairs of neighboring atoms, not constituting a cyclic bond. Finally for the pair  $\{x, a\}$ , we find an atom  $b$  such that  $d(x, b) = d(a, b) = D$ . We suspect that atoms  $x, a$ , and  $b$  lie on a hitherto undiscovered ring. However, we have to rule out the case where  $b$  is on an acyclic chain connecting two rings. To eliminate this possibility, we look for a pair of neighbors of atom  $b$ , such that one of the neighbors,  $bnbr1$ , is at a distance  $d - 1$  from  $x$ , and the other neighbor,  $bnbr2$ , is at a distance  $d - 1$  from atom  $a$ . If this pair is found, we have ruled out the case wherein atom  $z$  lies on an acyclic chain. If  $d = 2$ , we have found the ring. It consists of the atoms  $x, bnbr1$  which is  $y, z'$ ,  $bnbr2$  which is  $z'$ , and  $a$ .

If  $d$  is larger than 2, we have the general situation depicted in Figure 14, where we find an odd-membered ring of size  $2D$

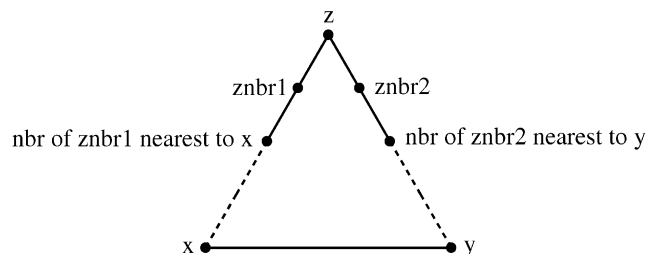
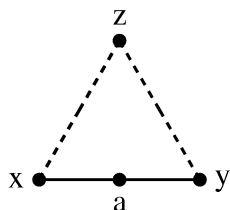


Figure 14.

+ 1. Atoms  $x$  and  $y$  are known to be neighbors of each other and their distance from  $z$  is  $D$  in both cases. Here we still need to collect the  $D - 1$  intermediate ring atoms intervening between  $znbr1$  and  $x$  and those intervening between  $znbr2$  and  $y$ . Thus, we find the neighbor of  $znbr1$  nearest to  $x$  and add it to our path. Then we find the neighbor of this atom nearest to  $x$  and add it to our path, and so on until we reach  $x$ . Similarly, we

find the neighbor of *znbr2* nearest to *y* and then the neighbor of that atom nearest to *y*, etc. until we reach *y*. We have thus collected all the atoms of the ring. They are *x*, ..., *znbr1*, *z*, *znbr2*, ... *y*. The chains represented by ... are of length  $D - 2$ .

**4.2. Outline of the Method for Finding Even-Sized Cycles.** The even-sized rings are found somewhat similarly. Here, instead of beginning with an edge, i.e., a pair of atoms, we start with a triplet of atoms, two atoms *x* and *y*, separated by a graph distance of two edges, and their common neighbor, atom *a* (see Figure 15). Next we proceed as with the odd-sized rings; i.e.,



**Figure 15.**

we seek an atom *z*, not *a*, that is equidistant from *x* and *y*. This distance is  $D$ . As in the odd-numbered ring case, there must be  $2D + 1$  atoms on the path from *x* to *z* plus the path from *z* to *y*. The paths in question must be the shortest possible paths. In this even-numbered ring case we must add atom *a* to complete the ring, giving this ring a total size of  $2D + 2$ . The initial value of  $D$  is 1, as 4 is the size of the smallest possible even-numbered ring. The  $D - 1$  atoms intervening between *x* and *z*, as well as those between *z* and *y*, can now easily be collected, using the nearest neighbor principle.

Here again, the ring in question is automatically one of the smallest set of smallest rings because we recognize the rings from the smallest to the largest. For example, we would find two six-membered rings in naphthalene and this exhausts the number of rings in the smallest set that includes all the cyclic atoms. We would thus never "see" the ten-membered ring. If the naphthalene was connected to another ring system with a twelve-membered ring, we still could not "see" the ten-membered ring because our algorithm forbids under most conditions the reuse of a known cyclic edge to begin a new ring. Here we require that at least one of the pair (the edge between *a* and *x*, the edge between *a* and *y*) must not appear in any previously discovered ring.

Thus we find the other atoms of the ring that consists of atoms *a*, *x*, ..., *z*, ..., *y*. We replace ... by atoms when we find the neighbor of *x* that is nearest to *z*, and the neighbor of that atom which is nearest to *z*, etc.

**4.3. Detailed Steps in the Procedure for Finding the Cycles.** A. Calculate the number of rings, in the smallest set of smallest rings. Our symbols are: *n*rings, the number of rings in the smallest set; *n*edges, the number of edges in the molecule; *n*vertexes, the number of vertexes in the molecular graph. The chemical meanings are, respectively, the number of rings in this smallest set, the number of bonds (single, double, triple or whatever), and the number of atoms in the molecule. A universally known equation is that  $n$ rings =  $n$ edges + 1 - *n*vertexes. If *n*rings is zero we exit.

B. Prune away all atoms of local degree 1, i.e., terminal atoms. Repeat this process until there is nothing left to prune. This removes all acyclic structures attached to rings except chains connecting rings. If *n*rings is unity, then we do procedure C. In a multicyclic molecule there is the possibility of a ring assembly, i.e., some rings connected by acyclic chains.

C. When there is a single ring in the molecule, start with an arbitrary unpruned atom and make a list beginning with the

atom, then one of its neighbors, next a neighbor of this neighbor that is not yet on the list and so on, until all the unpruned atoms are on the list. This list comprises the atoms of the single ring in order of traversal. We then exit.

D. When there are multiple rings, we do the following steps:

i. Calculate all the distances,  $d(i,j)$  as above.

ii. Find the three-membered rings as follows: Consider all pairs of unpruned atoms *x* and *y* in the molecule, such that *y* is greater than *x*. For each such pair look for a neighbor of *x* among the neighbors of *y*. If we have found such an atom, *z*, then the atoms *x*, *y*, and *z* constitute a three-membered ring. In this algorithm, whenever we find a ring that has just one atom of degree 3 and no other atoms of degree greater than 2, then we prune away the entire ring. We next examine all the atoms of degree 2 that intervene between an atom *a* and the nearest atom of degree higher than 2, to be called *b*. Having found such an atom, we then delete atom *a* and all the atoms on the chain between *a* and *b*, not including atom *b*.

iii. Set an integer variable, *currentSize*, equal to 3.

iv. Increment *currentSize* by 1. If *currentSize* is greater than the number of unpruned atoms in the molecule, then we have the case of multiple embedded rings, so we go to step xii. If *currentSize* is even, then we go to step ix for the even size ring procedure. Otherwise, we go to step v for the odd size ring procedure.

v. Begin the odd size procedure. Set  $D = (\text{currentSize} - 1)/2$ . Scan for a pair of atoms *x* and *z*, not previously examined in this step, which are separated by a distance of  $D$ . When there is no such pair remaining but we have found fewer rings than the calculated number rings, then return to step iv, deleting the requirement that the starting edge should not be in a ring that has previously been found. This is the case where we have multiple surrounded rings.

vi. Scan the neighbors of *x*. If one of them, to be called *y*, is also at a distance of  $D$  from *z* and the edge between *x* and *y* is not known to be cyclic, then we go to step vii. Otherwise, we return to step v.

vii. Because *y* is such that  $d(x,z) = d(y,z) = D$ , we look at the neighbors of *z* to see if there is one that is  $D - 1$  edges away from *x* and another neighbor that is  $D - 1$  edges away from *y*. If such two neighbors of atom *z* are found, then we have located a ring. When  $D$  is 2, we have a five-membered ring. This is the set of atoms  $\{x, y, \text{znbr2}, z, \text{znbr1}\}$ . More generally, the ring found has length  $2D + 1$ . It will initially contain the five atoms *x*, *y*, *z*, the neighbor of *z* that is  $D - 1$  edges from *x*, namely *znbr1*, and *znbr2*, the neighbor of *z* which is  $D - 1$  edges away from *y*. Then we generate the path from *x* to *znbr1*, without any blind traversal, by finding the nearest neighbor of *x* to *znbr1* and then the nearest neighbor of that atom to *znbr1* and so on, until the latest found is a neighbor of *znbr1*. By the same method of successive nearest neighbor location, we complete the last part of the ring, the string of atoms between *y* and *znbr2*. When we have found all these atoms, we have found a ring of the smallest set having length  $2D + 1$  or *currentSize*. If the ring has only one atom that in the pruned structure is of degree 3 (i.e., it has three neighbors) and no other atoms that have a degree higher than 2, then we prune away the ring and any attached acyclic chain leading to another ring. Return to the beginning of step vii.

viii. Having finished the scan over the unpruned atoms, we have found all rings of length *currentSize* and go back to step iv.

ix. Scan all the unpruned atoms A. For each such atom we scan its unpruned neighbors to find a pair of atoms *x* and *y* such that this pair has not previously been examined in this

step, with the current values of *currentSize* and *A*. Either the edge between *A* and *x* or the edge between *A* and *y* must not be in a previously discovered ring. If there is no such pair *x* and *y* for any unpruned atom *A*, then go back to step viii.

x. Select an atom *z*, other than *a*, *x*, and *y*, which has not previously been used in this step with the current values of *currentSize*, *a*, *x* and *y*. Atom *z* must be equidistant from *x* and *y* and the distance  $d(x,z) = d(y,z) = D$ . If there is no such *z*, then go back to step ix.

xi. If *z* is not equidistant from *x* and *y*, the distance being *D*, then go back to step ix. Next find *znbr1* and *znbr2*, being a distance of  $D - 1$  edges away from, respectively, *x* and *y*. We now have discovered an even-membered ring of length  $2D + 2$ . If we have finished the scan that began in step ix, and the number of rings found is less than the calculated smallest number of smallest rings, then return to step iv.

xii. The Embedded Ring Case. Usually at this point we have found all the rings of the molecule that are in the smallest set of smallest rings. Consequently, all atoms not found to be cyclic and not pruned must lie on chains connecting rings. An exception is the case where the graph has multiple surrounded (embedded) rings which are such that each of their edges are part of another ring. An example is shown in Figure 16.

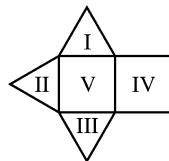


Figure 16.

Ring V is embedded in the other rings. It has no unique edge. Assume that we have first found rings I, II, III, and IV. At this point there are no acyclic edges left with which to begin a new ring. There remains ring V, which is an embedded ring of the smallest set. All of its edges are in other rings. The situation is detected by the program from the fact that the number of rings found is less than the calculated size of the smallest set of smallest rings and yet there are no more edges still considered to be acyclic. In this case we drop the requirement that the starting edge must not be a part of a ring that has already been discovered and resume the discovery process. This algorithm is thus slower in finding embedded rings than it is in finding nonembedded rings.

**4.4. Some Differences of This Algorithm from That of Figueras.** The breadth of the first algorithm of Figueras<sup>10</sup> depends on the generation of paths, and the examination of the intersections of the paths, when the trail down one path reaches an atom that itself has a path leading up to it. A ring has been found when there is only one atom that lies on both paths. In his words, the intersection of the paths is a singleton. In our algorithm, e.g., in the odd size rings situation, we must first spend a considerable time searching for a set of three atoms as described above, i.e., an adjacent pair and a lone atom that is equidistant from the that pair. Then we must seek a pair of neighbors of the lone atom such that each are the same distance from one of the starting pair of atoms. However, unless there is an intervening acyclic chain, we are now sure that a new ring has been found. Essentially, like the algorithm of Figueras, our algorithm looks ahead and does not find meeting paths that have larger than a singleton intersection. "Looking ahead" also takes time, as we scan for a suitable pair or triple. Even after optimizing our code, the algorithm of Figueras is about twice as fast as ours for typical cases. We hope in the future to explore

the properties of an algorithm which will combine the Figueras algorithm and ours.

**4.5. Discussion: Certain Rings Outside the Smallest Set of Smallest Rings.** Any practical algorithm for finding the synthetically important rings of a molecule has to find other rings besides those in the smallest set of smallest rings, e.g., all the rings in a bridged system. When two rings have a chain of more than two atoms in common, this chain is recognized as a bridge. The two terminal atoms in this chain are recognized as bridgeheads. There is a third ring that includes the two rings and the bridgehead atoms minus the other atoms of the bridge. The two smallest of the three rings are sufficient to define the system and the smallest number of smallest rings in a bridge system is calculated to be 2. Consequently, the third ring is not in this smallest set. Our ring-finding program, which is part of an organic synthesis design system, finds such additional rings after finding the smallest set of smallest rings and subsequently recognizing the presence of bridge(s).

It is also true that in the case of embedded rings some molecular graphs do not have a unique smallest set of the smallest rings. In a complicated example cited by Lynch,<sup>8</sup> there are many three-membered rings and four four-membered rings. The four four-membered rings are not equivalent. The cardinality of the smallest set of smallest rings of the graph is one more than the number of three-membered rings. Hence an algorithm for finding the smallest set of smallest rings would find all the three-membered rings but just one of the four-membered rings, the final result therefore being ambiguous. For an infallible retrieval system, one must obtain all of the four-membered rings in this case. Lynch refers to this as the extended set of smallest rings. Because such molecules are not practically synthesizable, and in light of expected costs and benefits, we therefore confine our attention to the smallest set of smallest rings, accepting nonuniqueness in these extremely rare cases, thereby simplifying the algorithm for synthetically practical cases.

**4.6. Efficiency Considerations.** The scheme adapts to parallel execution. For each connected pair of atoms a separate processor can examine a particular one of the remaining atoms to see if it is equidistant from the pair with the particular distance *D*. As for efficiency, throughout we are looking for an atom *X* and for another atom *Z* at a certain distance from *X*. Evidently, the algorithm has complexity at least of order  $O(n^2)$ . In addition, in the odd-membered ring case we search the neighbors of atom *X* for an atom *Y* that is also at this distance from *Z*. So, for an odd-membered ring the algorithm is of order  $O(n^2)$  and, in fact, will increase with  $L(n^2)$ , where *L* is the average local degree of the vertexes of the graph. In chemical terms, *L* is the average number of atoms bonded to an atom, the average number of nearest neighbors. For an even-numbered ring the complexity is also  $O(n^2)$  and part of the proportionality constant is  $L(L - 1)$ . The proportionality constant will be larger for the even-sized rings because we are required to find two atoms among the neighbors of *X* such that they are both a distance of *D* from *Z*.

Figueras's algorithm<sup>10</sup> finds the later rings more rapidly than it finds the first rings. This is because of the deletion of certain cyclic atoms. We use this feature only for rings that cannot share atoms with other rings because all but one of the atoms in the ring are of degree 2 and the remaining atom has degree 3. In this case there can be an acyclic chain connecting this unique atom to another ring. If such a chain exists, it too is deleted.

We have used the present algorithm as a tool inside of a synthesis design program for the last fourteen years. Millions of chemical structures have been examined with our implementation of this algorithm. We have not detected failure to

recognize the correct rings. Such failures would from time to time give rise to chemical mistakes. For example, if four of the atoms in a naphthalene structure were considered to be in a ten-membered ring but not in a six-membered ring, then relationships such as meta and para, which presume membership in a six-membered ring, would not be noted properly and the situation would produce errors.

### 5. Application of Atom–Atom Distances to the Perception of Molecular Substructures of Interest in Organic Synthesis

Synthetically interesting substructures include those that are the produced substructure of some known synthetic reaction. For example, we cite the case of molecules with two ketone carbonyl groups, C=O. When the distance between the two carbonyl carbons is 1, i.e., they are adjacent, we have the associated reactions of 1,2-diketones. Amplifying this, we can make a table as follows:

distance	name of substructure	a corresponding reaction
1	$\alpha$ -diketone	RuO <sub>2</sub> oxidation of an alkyne
2	$\beta$ -diketone	$\gamma$ -alkylation with NaNH <sub>2</sub> + RX
3	$\gamma$ -diketone	ketone + $\alpha$ -bromoketone, via enamine
4	$\delta$ -diketone	$\alpha,\beta$ unsaturated ketone + enol silyl ether, Ti catalyst
5	1,6 diketone	ring opening of 1,2-diacyl cyclobutane

There are often many reactions corresponding to a single produced substructure. We cite the above merely as examples. It is evident that a recognizing apparatus for produced substructures of known reactions must make extensive use of the distance properties of functional atoms and other significant atoms of the molecule (such as ring junctions, etc.).

**Acknowledgment.** We gratefully acknowledge the support of this research by Iwao Dohgane and Hiroshi Yamachika, the directors of the Sumitomo Chemical Organic Synthesis Research Laboratory.

### References and Notes

- (1) Ray, L. C.; Kirsch, R. A. *Science* **1957**, *126*, 814.
- (2) Sussenguth, E. H. *J. Chem. Doc.* **1965**, *5*, 36.
- (3) Morgan, H. J. *J. Chem. Doc.* **1965**, *5*, 107.
- (4) Cieplak, T.; Wisniewski, J. L. *Molecules* **2001**, *6*, 915.
- (5) Dohgane, I.; Takabatake, T.; Bersohn, M. *Recl. Trav. Chim. Pays-Bas* **1992**, *111*, 291.
- (6) West, D. B. *Introduction to Graph Theory*; Prentice Hall: Englewood Cliffs, NJ, 1996 (or any other standard text).
- (7) Downs, G. M.; Gilley, V. J.; Holliday, J. D.; Lynch, M.; Wilkins, C. L. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 172, 186.
- (8) Downs, G. M.; Gilley, V. J.; Holliday, J. D.; Lynch, M.; Wilkins, C. L. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 187.
- (9) Downs, G. M.; Gilley, V. J.; Holliday, J. D.; Lynch, M.; Wilkins, C. L. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 207.
- (10) J. Figueras *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 986.
- (11) Balducci, R.; Pearlman, R. S. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 822.