Identification of Symmetries in Molecules and Complexes

Wei Chen,[†] Jing Huang,[‡] and Michael K. Gilson*[,†,‡]

Center for Advanced Research in Biotechnology, University of Maryland Biotechnology Institute, 9600 Gudelsky Drive, Rockville, Maryland 20850, and VeraChem LLC, 20010 Century Boulevard, Suite 102, Germantown, Maryland 20874-1118

Received January 15, 2004

An algorithm is presented that quickly detects local and global symmetries of single molecules and complexes. Based upon the Morgan Naming Algorithm, the algorithm involves traversing the molecule from a starting atom and building up a molecule name based upon the names of the atoms encountered along the traversal. Additional molecule names are generated from other starting atoms, and name-name matches are identified as corresponding to symmetry operations. A number of enhancements relative to prior methods yield increased efficiency and extended functionality. In particular, the present method detects not only global symmetries but also local symmetries associated with bond rotations as well as symmetries that are only apparent when alternate resonance forms are considered. Importantly, the present method works not only for single molecules but also for multimolecular complexes. As a consequence, it is well, and perhaps uniquely, suited to applications in supramolecular and host−guest chemistry. Applications include filtering out redundant conformations during conformational searching and free energy calculations; accelerating ligand−receptor docking calculations by reducing the sampling ranges of rotatable bonds linked to locally symmetric groups, such as phenyls; and automating the calculation of symmetry numbers for thermochemical applications.

INTRODUCTION

Molecular symmetry plays a role in diverse aspects of physical chemistry, including chemical bonding, thermochemistry,[1,2] and spectroscopy. Detection of symmetries is also important in molecular modeling. Thus, ligand−receptor docking calculations can save time if angular ranges are limited to avoid generating symmetry-related conformations; for example, phenyl rotations can be limited to a range of $0-\pi$ rather than $0-2\pi$ without any loss of information. More generally, in conformational searches, symmetry information is needed to eliminate repeat conformations from the output, and eliminating repeats during the actual search can avoid wasting time on repeat calculations. Eliminating repeat conformations is particularly important for algorithms that compute free energies as sums over low-energy conformations,[3−5] because error and computational costs both rise when a single conformation is counted more than once due to lack of recognition of a symmetry. Finally, algorithms that automatically construct compounds by linking chemical fragments can avoid inadvertently constructing the same compound twice by ensuring that two linking points related by a rotational symmetry are recognized as chemically equivalent. In all of these applications, it is useful to account not only for symmetries associated with the molecule as a whole (global symmetries) but also symmetries associated with portions of the molecule (local symmetries), such as the phenyl flip mentioned above. In addition, applications in supramolecular chemistry require a method that can detect symmetries not only for single molecules but also for multimolecular complexes.

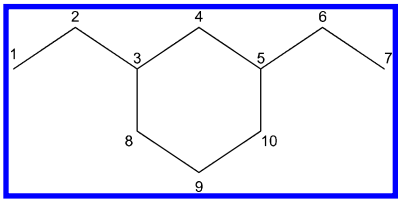


Figure 1. Topological versus 3D symmetry.

Two types of molecular symmetry are of interest.[6] A "topological" or "two-dimensional" (2D) symmetry is a conformation-independent mapping between groups of chemically equivalent atoms in a molecule or a complex, while a "three-dimensional" (3D) symmetry is a conformation-dependent mapping between chemically equivalent atoms that can be realized by a spatial rotation or reflection. For example, the molecule in Figure 1 possesses a topological symmetry that maps atoms (1,2,3,8) as a group to atoms (7,6,5,10), respectively, no matter what the conformation of the molecule may be. (If atoms 4 and 9 are considered to be included in the symmetry operation, they map to themselves.) On the other hand, a 3D symmetry that maps atoms (1,2,3,8) to (7,6,5,10) exists only if the conformation of the molecule is also symmetric, with an axis of rotational symmetry or a plane of reflection symmetry passing through atoms 4 and 9. Topological symmetries are particularly useful in computational molecular construction and conformational analysis, while 3D symmetries are important in, for example, spectroscopic and crystallographic studies.

Several algorithms for detecting the symmetries of three-dimensional objects have been described. Some rely on calculations of structural characteristics, such as moments

* Corresponding author phone: (301)738-6217; fax: (301)738-6255; e-mail: gilson@umbi.umd.edu.
[†] University of Maryland Biotechnology Institute.
[‡] VeraChem LLC.

10.1021/ci049966a CCC: $27.50
Published on Web 05/25/2004

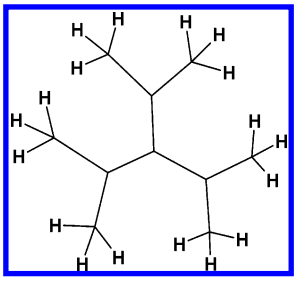


Figure 2. Compound with multiple symmetric branch points.

of inertia for molecules or the gradient orientation for polyhedrons.[7−9] However, these methods do not detect topological symmetries or local symmetries. In another approach, molecules are represented as graphs consisting of vertices (atoms) and edges (bonds), and symmetries are detected by testing for graph isomorphisms.[10−13] However, the graph isomorphism methods have, to our knowledge, addressed symmetry at the topological level only, and thus have not been adapted for detecting 3-D symmetries.

Recently, Ivanov and Schuurman have reported an algorithm that rapidly detects all global topological symmetries and 3D rotation and reflection symmetries for a range of test molecules.[6] The method is based upon the concept that detecting symmetry is closely related to testing the equivalence of different ways of labeling the atoms in the molecule.[15] In the algorithm, an atom of lowest bond connectivity is chosen as the starting point for a traversal of the molecular structure according to certain traversal rules, and a molecular "reference name" is generated based upon the resulting sequence of atoms and bonds. All alternative traversals that start from the same starting point and from additional starting points with the same bond connectivity are then used to generate additional molecular names. These additional names are compared with the reference name, and, if a perfect match is found, then the corresponding atoms in the two names must be related by a symmetry operation. With further analysis, the type and order of the symmetry can be determined. This method is conceptually appealing and has been implemented in fast computer code. However, it does not detect local symmetries or symmetries of multimolecular complexes. Also, it can miss symmetries in compounds with multiple resonance forms, because symmetry may not be evident unless the existence of alternate resonance forms is accounted for. Finally, the method is expected to become time-consuming and memory-intensive for molecules with many potentially symmetric branch points, such as that in Figure 2. Here, the 18 peripheral hydrogen atoms will be the starting points, and each traversal will encounter 3 branch points that are 2-fold symmetric and 5 branch points that are 3-fold symmetric. As a consequence, at least 34 992 names must be generated and compared.

The present paper describes an enhanced algorithm for detecting molecular symmetries that is inspired by the method of Ivanov and Schuurman. At the outset, alternate resonance forms are generated and averaged, ensuring that chemically equivalent atoms are treated as such. Then a set of appropriately chosen starting points is used to initiate molecular traversals that yield a single name per starting point. During the traversals, data are stored that permit reconstruction of all the traversals from a given starting point that would yield the same name. Note, however, that the method does not require storing all possible names associated with the starting point. The names associated with the various starting points are compared, and, if two names match, then the corresponding atom sequences must be related by a symmetry operation. Further analysis reveals the types and orders of the symmetries. This approach can markedly reduce the number of molecular names that must be generated and compared, relative to the prior naming approach. For example, only 18 names are required for the compound in Figure 2. In addition, efficiency is increased by assigning a highly descriptive but easily calculated code to each atom. This is useful because most branch points that are asymmetric can be determined to be asymmetric simply by comparing the codes of the first atoms along the branches; it is rarely necessary to compare the entire branches. Additional modifications accelerate the detection of 3D symmetries, and the present method is further equipped to detect local symmetries and the symmetries of multimolecular complexes. Thus, the method is both rapid and general.

METHODS

Overview. The purpose of the present algorithm is to detect the symmetries of a molecule or a multimolecular complex, i.e., to enumerate its automorphism group. The method is based upon a technique for generating a molecule name, where the name consists essentially of a sequence of atom names ordered according to a deterministic traversal of the molecule's atoms and bonds and does not depend on the initial ordering of the atoms presented to the algorithm. One name is generated from each of a specific set of starting atoms (see later in this section), and data are stored that enable reconstruction of the multiple traversal sequences that correspond to each molecule name. The names from the various starting atoms are compared with each other, and, if two or more names are identical, then all the associated traversal sequences from one starting atom are symmetry-related to all the corresponding traversal sequences from the other starting atom. This procedure identifies the topological symmetries; additional processing is used to determine which of the topological symmetries also correspond to a 3D symmetry.

The naming technique is such that each starting atom can yield only a single molecule name during detection of topological symmetries. Accomplishing this would be simple if the molecule were unbranched, because then each starting atom would trivially lead to a single obvious traversal. For most molecules, however, it is necessary to address the problem of branches. This is accomplished by applying a convention for ordering two branches linked to a common branch point. Initially, an attempt is made to distinguish the two branches based upon a descriptive atom code that is assigned to each atom in the molecule. If the atom codes of the first atoms along the two branches differ, they are ordered according to their codes and the traversal proceeds down the first branch and then returns to the branch point and continues down the second branch. If the atom codes are the same, then the order in which the potentially symmetric branches are traversed is recorded the first time the branch point is reached, and the same order is used for subsequent traversals in order to ensure that no symmetry is missed. If the two potentially symmetric branches prove to have identical subnames, then the order of their traversal

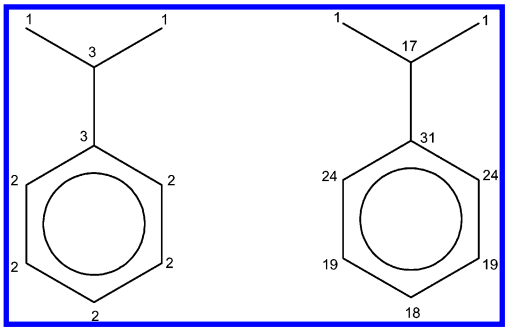


Figure 3. Connectivities (left) and extended connectivities (EC) (right) of cumin, with hydrogens omitted for simplicity.

does not affect the molecule name. In this case, the branch point is termed a "symmetry point", and the equivalence of the two branches indicates the existence of a local topological symmetry. Detection of 3D symmetries requires that conformational information associated with each atom be included in the molecule names, so that two molecule names match only if there is a conformational match, as previously described.[6]

Use of Atom Codes To Speed the Algorithm. There are a number of steps in the present algorithm in which it is necessary to decide whether two atoms could possibly be mapped into each other by a symmetry operation. If the atoms are obviously different, then the search for a symmetry can be stopped early, speeding the algorithm. As an example, if atoms were coded only by their elemental symbol, a methyl carbon and a carbonyl carbon would be nominally the same, and it would be impossible to determine that they are not equivalent under a symmetry operation without further examining their context in the molecule. However, if the codes for the same two atoms also included the number of other atoms to which they are bonded, the methyl code (C4) could immediately be distinguished from the carbonyl code (C3), so it would be clear that they are not equivalent under a symmetry operation and further structural analysis could be avoided.

Accordingly, each atom is here assigned an atom code that includes information not only about the atom itself but also about its location within the molecule. The code consists of the atom's elemental symbol, its extended connectivity (EC),[16,17] and its atomic partial charge. The atomic partial charges are computed with a rapid method described elsewhere,[18] and the EC values are calculated via the following steps: (1) Set the EC of all atoms to their connectivity—the number of other atoms to which they are bonded. (2) Count the number of different EC values in the molecule, N_{EC}. (3) Calculate a new EC value of each atom by summing the old EC values of its attachments, except that the EC of every atom with only one attachment remains fixed at a value of 1. (4) Determine the new value of N_{EC} and compare it with the old one. If the new value is greater than the old one, then go to step 3. If the new value is the same as the old one, the calculation is done, and the current EC values are the final ones. If the new value is less than the old one, the calculation is done, and the prior EC values are the final ones.

As an example, Figure 3 shows the connectivities and EC value of cumin. When it is necessary to put individual atoms in order, they are sorted first by element name, then by EC, and then by partial charge.

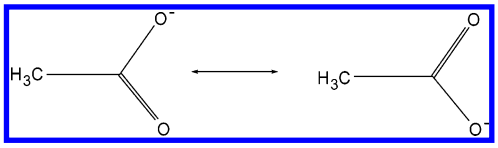


Figure 4. Resonance forms of acetate.

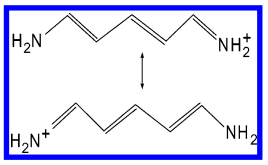


Figure 5. Resonance forms of a vinylogous cation.

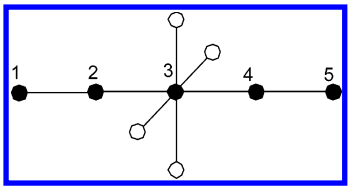


Figure 6. Diagram of a molecule whose starting atoms (unfilled) all lie in a plane normal to the paper. Some atom indices, i_{file}, are shown.

Treatment of Alternate Resonance Forms in the Naming Algorithm. Computer representations of molecules typically store only a single resonance form, but the electronic structures of many molecules are more accurately represented as averages over multiple resonance forms. For example, the two resonance forms of the acetate molecule in Figure 4 contribute equally to the molecule's electronic structure, and failure to recognize the equivalence of acetate's two oxygens would cause the molecule's obvious symmetry to be overlooked. Because acetate's two oxygens are close to each other, and a carboxylate is a standard chemical group, it would be possible to recognize this specific group and thus assign the two oxygens and their bonds as equivalent. However, a more general treatment of resonance forms is essential in more complex cases. For example, although the two nitrogens in the vinylogous compound in Figure 5 do not form a recognized chemical group, it is clear that they are chemically equivalent and that the molecule is symmetric. The present method solves this problem by using a recently developed algorithm[18] to identify alternate resonance forms and then assigning appropriate fractional bond orders. Note, too, that the partial atomic charges used here are computed with a method that accounts for alternate resonance forms[18] and hence assigns equal partial charges to chemically equivalent atoms.

Selection of Starting Points. If one used every atom in the molecule as a starting atom, one could be sure of detecting all name−name equivalences and hence all symmetries. However, this exhaustive approach is rarely necessary, based upon the following reasoning. Every global symmetry operation maps every atom in the molecule either to itself or to another atom with the same atom code. (Atom codes are defined in a previous subsection.) This holds true for the starting atoms also, and thus indicates that, so long as no symmetry operation maps any starting atom to itself, then the set of starting atoms needs to include only atoms with a single atom code. We now consider the case where every starting atom is mapped to itself by a symmetry operation, as illustrated in Figure 6. (The case where only some of the starting atoms are mapped to themselves is

considered a little later in this section.) If the white atoms are used as the starting points, then the 4-fold rotational symmetry will be detected by the equivalence of the 4 molecular names, but the global 2-fold reflection symmetry which maps each white atom into itself will be missed. If atoms 1 and 5 are used as starting points, the global 4-fold rotation will be missed, but the global 2-fold reflection which maps each white atom into itself will be detected.

If we are interested only in topological symmetries, then missing these global symmetry operations is not problematic because the symmetry-related atom−atom interchanges are still detected as local symmetry operations and recognizing that the starting atoms map to themselves adds no information of practical value in this context. For example, in Figure 6, all the starting atoms (open circles) map to themselves in a reflection symmetry through the reflection plane defined by the starting atoms. The only meaningful interchange that occurs is that of atoms 1 and 2 with atoms 5 and 4, respectively, and this interchange is detected as a local symmetry no matter which of the starting points is used, due to the equivalence of the 2−1 and the 4−5 branches that originate from atom 3. Thus, the atom interchange that corresponds to the reflection symmetry is detected and the fact that the associated symmetry is categorized as local rather than global has no influence on the practical applications envisioned in the Introduction.

If we are interested in 3D symmetries, it is necessary and straightforward to check for the possibility of a symmetry that maps all starting atoms to themselves. The present procedure is based upon the facts that a reflection symmetry can map all starting atoms to themselves only if the starting atoms are coplanar, and a rotation symmetry can map all starting atoms to themselves only if the starting atoms are not only coplanar but also colinear. To begin, the entire molecule is checked for linearity and planarity. If the entire molecule is linear (to within a user-specified tolerance), then any set of starting atoms will be colinear, no nontrivial symmetries will be missed due to the colinearity of the starting atoms, and no further steps are needed. Similarly, if the entire molecule is planar and the starting points are not colinear, then no additional starting points are needed. However, if the molecule is planar and the initial starting points are colinear, or if the molecule is nonplanar and the initial starting points are coplanar, then additional starting points are needed. In these cases, the second-largest atom code group is added to the original set of starting atoms. If these new starting points are coplanar with the original starting points (or colinear, for a planar molecule), then the next atom code group is tried instead, and so on, until the expanded set of starting points is noncoplanar for a nonplanar molecule or noncolinear for a planar molecule. Rarely, it may be necessary to add more than one atom code group to the starting set. It is perhaps worth noting that, for a nonplanar molecule, at least 4 starting points are always required to ensure detection of all 3D symmetries.

Finally, we consider the case in which some, but not all, starting atoms are mapped to themselves by a symmetry operation. In this case, the symmetry will be detected by the mapping of the other starting atoms into each other. For example, if the atom code of atoms 2 and 4 in Figure 6 happened to match the atom code of the white atoms and they were all therefore used as starting atoms, then both the 2-fold reflection and the 4-fold rotational symmetries would be detected as global symmetries. Therefore, to maximize the number of topological global symmetries that are correctly categorized as global, rather than local, we use the largest atom code group for the starting atoms. If two atom code groups are equal in number, we use the one that has a higher rank, as described in the section on atom codes.

Molecule Names and Traversal Sequences. After atom codes have been assigned and starting points have been selected, the next step is to traverse the molecule from each starting point, generating molecule names and information on the atom sequence(s) corresponding to each name. Here, an atom sequence is a sequence of the integer indices i_{file} assigned to the atoms in the input file, ordered according to the order of the traversal. For topological symmetry-detection, the name of the molecule is a linear concatenation of atom names, where the atoms are listed according to the order of the traversal and each atom name is constructed as follows: elemental symbol//<stereo parity code>//([bond order]//sequence number of connected ancestor)$_n$, where // indicates concatenation, ()$_n$ indicates iterations over the n atoms already listed in the name to which the atom of interest is bonded, and angle brackets indicate that the parity is included only for chiral centers. The value of n is always 0 for the starting point because no other atoms are already listed in the name, and $n > 1$ only if the atom of interest forms part of a ring. Aromatic bonds are tagged by assigning a bond order of 6, and noninteger bond orders are allowed for molecules with multiple resonance forms. For example, the CO bonds of a carboxylate are assigned a bond order of 1.5.

The stereo parity code is included only for chiral atoms. If R and S specifications are available, the stereo parity code may be set to 1 or 2 for R or S, respectively. In the more common case where parity is specified in an MDL SDfile via an atom stereoparity code, stereowedge notation, or an initial 3D conformation,[19] a well-defined parity code can still be assigned to an atom by sorting its four substituents according to the atom code of the first atom of each substituent and, if these are identical, by sorting the subnames of its substituents as described in traversing the molecular structure, and then assigning a parity of 1 or 2 by the MDL method.[19]

For 3D naming and symmetry detection, each atom name is supplemented by adding a suffix consisting of the value of a dihedral angle defined with respect to 3 atoms already listed in the name. Thus, the atom code is as follows: elemental symbol//([bond order]//sequence number of connected ancestor)$_n$//[dihedral angle// defining atoms]. Clearly, the first 3 atom names cannot include dihedral angles. For each subsequent atom n, all bond-torsions involving 3 other atoms already in the name are identified and ranked according to the order in which the 3 defining atoms appear in the name. The torsion angle is then computed with respect to the top-ranked group of 3 atoms. If no such torsion angle exists, then an improper dihedral angle must exist and it is used instead. As previously pointed out,[6] 3D names could be extended to include bond lengths and bond angles, but this would add little additional information because these degrees of freedom are relatively rigid.

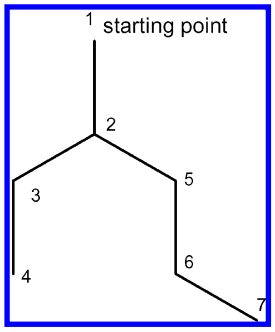


Figure 7. Figure of a molecule, along with its the integer indices i_{file}, that contains an asymmetric branch point (atom 2).

For example, if a hydrogen is used as the starting atom, then the topological molecule name of acetylene would be

H C [1] 1 C[3]2 H[1]3

and the 3D name of the expected linear conformation is

H C[1]1 C[3]2 H[1]3 [180 3 2 1]

If the input file happens to list the two carbons followed by the two hydrogens ordered by the carbons to which they are bonded, the two starting hydrogen atoms would generate two different traversal sequences, (3, 1, 2, 4) and (4, 2, 1, 3), but the same molecular name, indicating a symmetry between corresponding atoms in the two sequences.

For an example of a branched, asymmetric molecule, consider the compound illustrated in Figure 7, where hydrogens are neglected for simplicity, the starting atom is marked, and the traversal visits the shorter branch first. Here, the topological name is

C C[1]1 C[1]2 C[1]3 C[1]2 C[1]5 C[1]6

and the 3D name might be

C C[1]1 C[1]2 C[1]3[120 3 2 1] C[1]2[60 2 3 4] C[1]5[115 5 2 1] C[1]6[45 6 5 2]

Using the other carbons as starting points will clearly yield different names, so the molecule possesses no symmetry.

For an example of a molecule with a symmetric ring, consider toluene, neglecting hydrogens for simplicity and using the methyl carbon as the starting point. Here the topological name is

C C[1]1 C[6]2 C[6]3 C[6]4 C[6]5 C[6]6[6]2

where the last atom name lists two bonds as described earlier in this section. In this case, the ring carbon bound to the methyl has two branches, which correspond to clockwise or counterclockwise traversals of the ring. Since the two directions yield different atom sequences but the same molecular name, the branch point is also a symmetry point.

Traversing the Molecular Structure. The molecular structure is traversed according to a recursive algorithm. To begin, one of the starting atoms is designated the current atom (CA). It is added to the molecular name, and its index i_{file} is added to the associated atom traversal sequence. If only one other atom is bonded to the CA, the traversal proceeds to that atom, which is designated the new CA and is accordingly added to the molecular name and the sequence. If the CA is bonded to multiple other atoms, then the atom

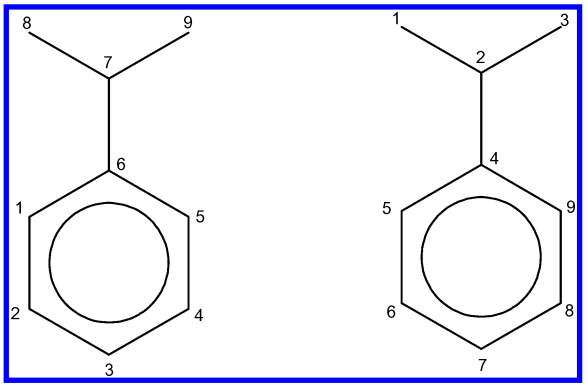


Figure 8. Diagram of traversal of cumin, with hydrogens omitted for simplicity. Left: atom sequence numbers from the input file (i_{file}). Right: order of traversal when atom 8 is used as the starting point. In a second traversal, atom 9 would be used as the starting point. In both cases, atom 7 is a branch point and atom 6 is a symmetry point.

codes of the branches are sorted as described above, and, if there is no tie for the highest atom code, then the traversal proceeds via the atom with the highest ranked atom code. The part of a molecule name generated starting at a branch point is called a subname for the branch point, and this subname is added to the molecular name when it is generated. If there is a tie for the highest atom code, then the branch may or may not be a symmetry point. If it is not a symmetry point, then the order in which its branches are visited will affect the resulting molecular name, and it is necessary to ensure that the same sequence of subnames is obtained if this branch point is reached later from a different starting point. Therefore, the first time this point is reached from say starting point A, the branches are traversed in an arbitrary order and the order is recorded for later reference. When the same branch point is reached again from a second starting point B, the branches are traversed in the same order, except that the branch leading to the original starting point A is interchanged with the branch leading to starting point B. It is also stored so that the subnames associated with a given branch point can be compared with each other; if two names are identical, the branch point is considered a symmetry point and a local symmetry is considered to exist.

Note that if the traversal of one branch returns to the same branch point, then the two branches are involved in a ring and only one of the subnames associated with the ring should be added to the overall molecular name so that the atoms will not be included twice. Nonetheless, if the first atom along one ring branch has the same atom code as the first atom along the other ring branch, then both names must be generated and ranked. If the resulting names differ, the higher ranked one is used, along with its sequence. If the resulting names are identical, the branch point is a symmetry point; the ring's subname is added to the molecule name, and sequence information is stored for both traversal directions. For example, in Figure 8, when atom 6 is traversed from atom 7, it is a branch point in which traversal of one branch leads to the other branch. Atom 6 also is a symmetry point for traversals starting from atoms 8 or 9. It is worth noting that an atom may be a symmetric branch point when approached along one bond but asymmetric when approached along another; hence each symmetry point is associated with a specific starting atom.

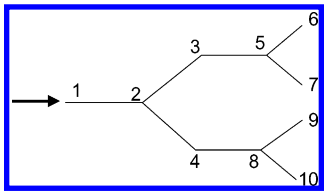


Figure 9. Molecular fragment with successive symmetric branchings. Traversal is considered to start at atom 1 due to a starting point elsewhere. Hydrogens are omitted for simplicity.

This procedure iterates until no new atoms are available along the growing sequence. At this point, there may still be unvisited branches at branch points for which the atom code ranking sufficed to establish the branch to be traversed first. The traversal therefore returns to the most recently traversed branch point that still has an unvisited attachment and continues in the same manner. This process is repeated until every atom in the molecule has been added to the molecular name and all the requisite sequence information has been generated.

Figure 8 illustrates the naming algorithm using cumin as an example, where atom $i_{file} = 8$ is used as the starting atom. The traversal (right) starts with atom 8 and generates the following molecular name, no matter which branch is selected from the symmetry point atom 6:

C C[1]1 C[1]2 C[1]2 C[6]4 C[6]5 C[6]6 C[6]7 C[6]4[6]8

Because atom 6 is a symmetry point, two sequences correspond to this name: 8 7 9 6 1 2 3 4 5 and 8 7 9 6 5 4 3 2 1.

Figure 9 illustrates a case that is more complex because it involves two layers of symmetric branch points. (Again, hydrogens are omitted for simplicity.) If the moiety shown is approached from atom 1, due to a starting point elsewhere in the molecule, the subname of the moiety will be

C C[1]1 C[1]2 C[1]3 C[1]5 C[1]5 C[1]2 C[1]4 C[1]8 C[1]8

This name corresponds to the following traversal sequences, where the symmetry points detected during the traversal are shown in bold font:

1	**2**	3	**5**	6	7	4	**8**	9	10
1	**2**	3	**5**	7	6	4	**8**	9	10
1	**2**	3	**5**	6	7	4	**8**	10	9
1	**2**	3	**5**	7	6	4	**8**	10	9
1	**2**	4	**8**	9	10	3	**5**	6	7
1	**2**	4	**8**	10	9	3	**5**	6	7
1	**2**	4	**8**	9	10	3	**5**	7	6
1	**2**	4	**8**	10	9	3	**5**	7	6

These alternate traversals are stored in a more compact form by storing each symmetry point, together with the branch sequences downstream from it up until any subsequent symmetry point:

2: (3 **5**) (4 **8**)

5: (6) (7)

8: (9) (10)

This information enables the full set of alternate traversals to be reconstructed as needed.

Detection of Global Topological Symmetries. The procedures described above are used to generate a molecule name and corresponding atom sequence data for each starting atom, where the atom sequence data include information permitting alternate atom sequences that give the same molecule name to be constructed for each symmetry point. A global topological symmetry may exist when the names from two starting atoms are identical; in this case, every atom sequence that corresponds to the first name is related by a symmetry operation to every sequence that corresponds to the second name. However, it may be argued that the identity of the two names is not a sufficient condition for existence of a global topological symmetry, as now discussed.

Consider the compound on the left in Figure 10. If the two lower chlorines are used as starting points, then two identical names will be generated and it might therefore be concluded that the molecule possesses a 2-fold topological symmetry. However, the CHBrCl group does not share the 2-fold symmetry of the ring and its immediate substituents, so one may legitimately question whether a topological symmetry exists. It is not immediately clear what criterion should be used to answer this question, but one approach is to base this assessment on whether atoms that are supposedly related by symmetry, such as atoms 1 and 5, are in equivalent molecular environments. This approach is of practical relevance in problems of molecular construction, because it bears on whether adding a given fragment to atom 1 or atom 5 will generate the same compound. Intuitively, the positions seem equivalent because the CHBrCl group can rotate around the marked bond, so positions 1 and 5 will be chemically equivalent over time. On the other hand, this equivalence is temperature dependent and in the topologically similar case shown in the middle of Figure 10, one might well conclude that positions 1 and 5 are distinct. To avoid missing distinct molecular attachment points in such a case, we prefer to regard the left two molecules in Figure 10 as lacking any global topological symmetry. It is therefore necessary to distinguish these cases from the one on the right, which is definitely symmetric.

This distinction is made by comparing the atom sequences of the matching molecular names and applying the criterion that every branched atom that maps to itself must be a symmetry point whose order is an integer multiple of the order of the putative global symmetry in order for a global topological symmetry to exist. In the present example, the atom sequences for the middle molecule might be

10/2/1/6/**7**/8/9/5/4/11/**3**/12

11/4/5/6/**7**/8/9/1/2/10/**3**/12

where branch atoms that map to themselves, atoms 7 and 3, are in bold font. Here, although atom 3 is a symmetric branch point which is discovered in the search for local symmetries (see next section), atom 7 is not, so the molecule is not considered to possess a global symmetry. The same sequences could be generated for the right-most molecule in the figure, but now atom 7 also is a symmetric branch point, so the molecule is considered to possess a global topological symmetry. Although this definition of global topological symmetry is appropriate for many applications, it is not

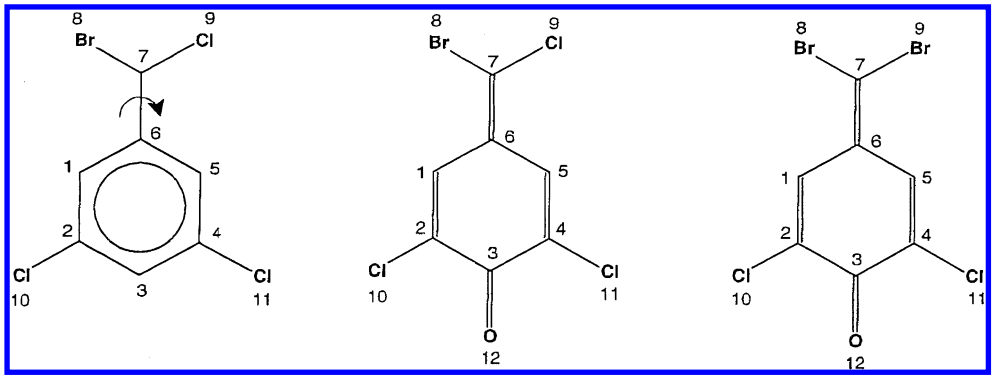


Figure 10. Compounds to illustrate an issue in the definition of topological symmetry. Atom indices, i_{file}, are shown.

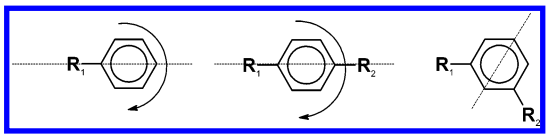


Figure 11. Local rotational symmetries are considered to be associated with two phenyl groups (left and center) but not a third (right).

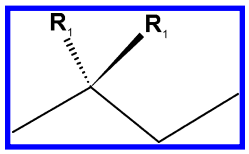


Figure 12. Local reflection symmetry associated with equivalent groups bonded to the same atom.

entirely adequate as a means of identifying potential 3D symmetries because the apparently asymmetric atoms—for example, the Br and Cl in the middle of Figure 10—may lie in the plane of a reflection symmetry. This possibility is specifically tested during detection of 3D symmetries, as noted below.

Detection of Local Topological Symmetries. In addition to global symmetries, we wish to detect two types of local symmetry that are of interest in molecular construction and conformational search. The first is a rotational symmetry about an axis that is colinear with a bond or bonds, as illustrated in Figure 11a,b, where a 180° rotation about the axis of local symmetry (dashed lines) regenerates the starting conformation. In a conformational search, there is no reason to report two conformations that differ only by such a rotation; indeed, there is no reason for the search range to be greater than 180°. In addition, the symmetry-related carbons of these phenyl groups represent chemically equivalent attachment points for molecular construction. This type of symmetry requires a symmetric branch point. The second local symmetry of interest here is a reflection symmetry involving groups attached to a single branch point atom, where the reflection plane passes through the branch point, as illustrated in Figure 12, where the two R_1 groups are related by a reflection symmetry passing through the branch point atom. This type of symmetry is of interest in conformational search algorithms that occasionally switch two groups attached to the same atom. In Figure 12, for example, it is useful to know that swapping the two R_1 groups does not result in a new conformation. This type of local symmetry also requires a symmetric branch point. Note that local rotational and reflectional symmetries that do not involve a symmetric branch point are of less interest for conformational analysis and molecular construction. For example, although the phenyl moiety in Figure 11c possesses multiple rotational axes of local symmetry (such as the dashed line) and planes of symmetry (not shown), these symmetry operations are of less interest here because they generate highly strained conformations. In addition, if the two R_1 groups in Figure 12 are linked to each other to form a ring, the local reflection symmetry is excluded because it is unlikely that a conformational search program will swap them.

Some of a molecule's symmetry points are found during the naming process used to detect global topological symmetries. However, although the traversals used to detect global symmetries do visit every atom and bond, detecting all symmetric branchings requires visiting every branched atom from every direction; that is, along each of its bonds. This point is illustrated for the molecule cumin in Figure 8. When atoms 8 and 9 are used as starting points, the symmetry of the phenyl ring is detected, with atom 6 as the symmetry point. However, the symmetry of the two extracyclic methyl groups is missed because the branched atom 7 is visited from only two of three possible directions. This problem is addressed as follows. Each atom with connectivity greater than 2 that has not already been approached along all its bonds during a previous traversal is identified. Such atoms are termed untested atoms, and the bonds that have not been used to approach them are termed their unused bonds. Each untested atom is then used as a starting atom for additional subnames, where the new traversals are initiated as if the atom is approached along one of its unused bonds. The resulting subnames for each untested atom are compared for matches and hence for new local symmetries. Note that the new traversals associated with one untested atom may arrive at other untested atoms along their unused bonds and thus may reduce the number of untested atoms. Therefore, with each new traversal, the set of untested atoms is reduced until all branched atoms have been fully tested and all local topological symmetries thereby identified.

Summary of the Cumin Example. For the simplified cumin molecule in Figure 8, with atoms 8 and 9 used as starting atoms, both starting points give the same molecule name with different atom sequences, depending upon the direction in which the ring is traversed. That is, the following atom sequences all yield the same name:

8	7	9	6	1	2	3	4	5
8	7	9	6	5	4	3	2	1
9	7	8	6	1	2	3	4	5
9	7	8	6	5	4	3	2	1

The detection of local symmetries described in the previous

section yields that atom 6 is a symmetry point where the corresponding local symmetry is the interchange of atoms (1 2 3 4) with atoms (4 3 2 1), effectively a 180 degree rotation of the phenyl group, and atom 7 is a symmetry point where the corresponding local symmetry is an interchange of atoms 8 and 9. Comparing the middle two sequences and recognizing that the only branch points that map to themselves, atoms 6 and 7, are both symmetry points (see previous paragraph) shows that there is a global 2-fold symmetry that maps the atoms in the second sequence to the corresponding atoms in the third sequence. This global symmetry represents a complete interchange of all the atoms on the left-hand side of the molecule with the corresponding atoms on the right.

Using Topological Symmetries To Eliminate Conformational Overlaps. As discussed in the Introduction, detecting molecular symmetries is an important step in eliminating repeated conformations during conformational search procedures. In the present algorithm, each topological symmetry operation is represented by an interchange of two groups of atoms that yields a symmetry-related conformation of the molecule. This information may be used to determine whether two conformations A and B of a molecule with N_{sym} symmetry operations are equivalent. This is done by applying all possible combinations of symmetrical atom interchanges to one of the conformations, say conformation A. The resulting symmetry-related conformations A_i are screened to remove mirror images, leaving only rotations, and the remaining rotation-related conformations A_i are used to compute the root-mean-square distance (RMSD) relative to conformation B. The true RMSD is the lowest one obtained, and conformations A and B are considered equivalent if the lowest RMSD falls below a user-specified tolerance.

3D Conformation Names. The discussion so far has focused on topological symmetries; i.e., on symmetries that exist independent of molecular conformation. This section describes further analyses to detect and characterize the 3D symmetries of a molecule in a given conformation. The method involves constructing names with 3D conformational information (see "Molecule Names and Traversal Sequences" above) for sequences that correspond to the same topological molecule name (see above); when two such conformation names match to within a specified tolerance, a 3D symmetry operation has been identified. Further analysis permits the corresponding symmetry elements to be determined. This procedure is carried out for both local and global topological symmetries.

Except for certain small molecules (see next paragraph) a conformation name is generated simply by calculating all the proper dihedral angles in the molecule and then ordering them according to the atom sequence for which a conformation name is required. For example, to generate the conformation name associated with the atom sequence 9 7 8 6 5 4 3 2 1 for cumin, the molecule's 12 dihedral angles are ordered as follows, where bold face type is used only to help the reader's eye distinguish between the integer quartets:

9-7-6-5 **9-7-6-1** 8-7-6-5 **8-7-6-1** 7-6-5-4 **1-6-5-4**
6-5-4-3 **5-4-3-2** 4-3-2-1 **3-2-1-6** 2-1-6-7 **2-1-6-5**.

Restricting attention entirely to proper dihedrals would make it impossible to detect symmetries in small molecules that have no proper dihedral angles, such as chloromethane and ammonia. Detecting symmetries in these molecules is not important in conformational search algorithms because their conformations do not vary significantly so long as their stereochemistries are considered fixed. However, in some applications it is helpful to be able to detect their symmetries so that their symmetry numbers may be computed. This problem is addressed here by omitting all dihedral information from molecule names of molecules with no proper dihedrals and thus not requiring dihedral matches as criteria for identifying 3D symmetries. However, all potential 3D symmetries detected by this method are explicitly checked by applying the symmetry operations and determining whether the transformed molecule coordinate can be overlayed on its initial coordinates to within a selected distance tolerance.

Comparison of 3D Conformation Names. In global symmetry detection, each starting atom can generate multiple traversal sequences. Each of these traversal sequences is associated with a conformation name, and detecting all the 3D symmetry operations that relate two starting atoms A and B might appear to require comparing the full set of conformation names from starting atom A with the full set from B, where an exact match corresponds to a global symmetry operation. In fact, however, the number of comparisons can be reduced by recognizing that multiple traversal sequences from a given starting point result only from symmetric branch points (i.e., symmetry points). Thus, in the absence of a branch point, there is obviously only one traversal sequence; and there is a definite order to the traversal of the branches of an asymmetric branch point, so an asymmetric branch point generates only one traversal sequences. Because a symmetric branch point of order n generates n different traversal sequences, an exhaustive comparison of names for starting points A and B would require $\prod_{i=1}^{m_A} n_{A,i} \prod_{i=1}^{m_B} n_{B,i}$ name−name comparisons, where m_A and m_B are the numbers of symmetry points with respect to the two starting points, and $n_{A,i}$ and $n_{B,i}$ are the orders of the corresponding symmetry points. In fact, however, it is only necessary to compare one conformation name from starting atom A to all the conformation names from starting atom B. This is because two traversals from a given starting atom differ only in the order in which the branches of the symmetry points are traversed. As a consequence, the different conformation names from starting atom A are just different orderings of the same set of dihedrals; and starting atom B, if it is related by symmetry to starting atom A, will have a full set of the same conformation names with all the same dihedral orderings. Therefore, all symmetry information can be obtained by comparing a single conformation name from starting atom A with all the conformations from starting atom B. This procedure reduces the number of name−name comparisons to $\prod_{i=1}^{m_A} n_{A,i}$. The size of this combinatorial problem can be further reduced by dividing each atom sequence into subsequences split at symmetry points. For instance, the cumin sequence 9 7 8 6 5 4 3 2 1 is divided at symmetry point 6 into 9 7 8 6 and 6 5 4 3 2 1. (See Figure 8.) Each subsequence is assigned a conformational subname, where dihedral angles are allocated based upon their central bond. For example, angle 9−7−6−5 is associated with the first name, and angle 7−6−5−4 is associated with the second name. Comparisons are then performed one subname at a

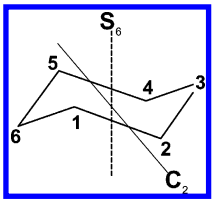


Figure 13. Chair conformation of cyclohexane (hydrogens omitted for simplicity), showing the axis of one of its 3 global 2-fold rotational symmetries, and the axis of its 6-fold improper rotational symmetry.

time. For example, if the first subname of the conformation name from starting atom A does not match the first subname of a conformational name from starting atom B, then the conformation name from B can be discarded immediately. By this procedure, the number of comparisons is reduced to a maximum of $\sum_{i=1}^{m_A} n_{A,i}$.

Cyclohexane, shorn of hydrogens, offers a simplified illustration with no symmetry points. When atoms 1 and 2 (Figure 13) are used as starting points, each yields two traversal sequences, and all four molecule names are the same:

C C[1]1 C[1]2 C[1]3 C[1]4 C[1]5[1]1

The traversal sequences and corresponding conformation names are as follows:

Starting atom 1

Traversal 1**:** 1 2 3 4 5 6 60 −60 60 −60 60 −60

Traversal 2: 1 6 5 4 3 2 −60 60 −60 60 −60 60

Starting atom 2

Traversal 3: 2 3 4 5 6 1 −60 60 −60 60 −60 60

Traversal 4: 2 1 6 5 4 3 60 −60 60 −60 60 −60

Comparison of the first conformation name from starting point 1 with the two conformation names from starting point 2 yields one exact match and one match with a sign-reversal. These matches correspond, respectively, to a rotational symmetry (C_2 axis in Figure 13) and to a rotation-reflection symmetry (S_6 axis in Figure 13).

Identification of 3D Symmetry Elements and Symmetry Orders. Three-dimensional molecular symmetries may be described in terms of two types of symmetry element: n-fold rotation (**C_n**; proper rotation) and n-fold rotation combined with reflection through a plane perpendicular to the rotation axis (**S_n**, or "improper rotation"). Note that reflection and inversion are equivalent to **S_1** and **S_2**, respectively, and so need not be treated separately. As previously discussed,[6] if two conformation names match exactly, then the corresponding atoms are related by a proper rotation, and if two names differ only in the sign of every dihedral angle (where π is considered equivalent to $-\pi$), then they are related by an improper rotation. If two equivalent conformation names consist of only 0 and π dihedral angles, then the molecule is planar and both proper and improper rotational symmetries exist. This subsection describes how the rotation axes and reflection planes associated with a molecule's symmetry elements can be computed.

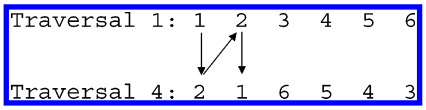


Figure 14. Counting the multiplicity of a 2-fold symmetry in cyclohexane.

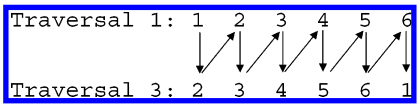


Figure 15. Counting the multiplicity of a 6-fold symmetry in cyclohexane.

First, the multiplicity, *n*, of each rotational symmetry is determined. For local symmetries, *n* equals the number of equivalent attachments of the associated symmetry point. For global symmetries, *n* is calculated by counting the number of atoms that are mapped to each other by the symmetry in question. This is accomplished as follows. The two atom sequences, here termed sequences **a** and **b**, whose name match led to detection of the symmetry, are aligned and the first atom in sequence **a** that is not mapped to itself in sequence **b** is identified. The corresponding atom in sequence **b** is noted, and the position of this atom in turn is found in sequence **a.** The symmetry-related atom in sequence **b** is then identified, as was done for the first atom. This procedure is iterated until the atom that started the procedure is reached in sequence **b**; the value of *m* is then the number of symmetry-mappings completed. Figure 14 illustrates the application of this procedure to the match of traversals 1 and 4 for cyclohexane (previous subsection). The alignment of the sequences indicates that atom 1 is related by this symmetry to atom 2; returning to atom 2 in traversal 1 and again looking across to traversal 4 shows that atom 2 is symmetry related to atom 1, indicating a 2-fold (rotational) symmetry, as illustrated in Figure 14. Similarly, the 6-fold rotation/reflection symmetry of this conformation of cyclohexane (**S_6** in Figure 13) is detected by comparing traversal 1 with a traversal 3, as shown in Figure 15.

Finally, the symmetry axes and planes are identified. For each symmetry, it is necessary first to define a reference point (RP). For a local symmetry, the RP is the associated symmetry point. For every global symmetry, the RP is the geometric center of the molecule, i.e., the average coordinates of all atoms, because all symmetry axes and planes pass through it. The detection of symmetry axes and planes is then the same for both local and global symmetries. For a proper rotation symmetry of order *n*, the *n* atom sequences that give identical molecule names (or, for a local symmetry, identical branch names) and conformation names are used to calculate a vector defining the rotation axis

$$\vec{V}_j = \sum_{i=1}^{m} (\vec{r}_{ji} - \vec{r}_{RP})$$

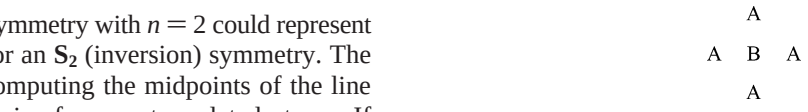
where $\mathbf{r}_{ji}$ represents the 3D coordinates of the *j*th atom in the *i*th atom sequence, and $\mathbf{r}_{RP}$ represents the coordinates of the reference point. Initially, the first atom in the atom sequences ($j = 1$) is used. However, if $|\mathbf{V_j}| = 0$ for $j = 1$, then the atoms in position 2 are used, and so on, until a nontrivial axis is identified. If $|\mathbf{V_j}| = 0$ for all j and $n > 2$, then the molecule must be planar and the symmetry axis is perpendicular to the plane of the molecule. In this case, $\mathbf{V_j}$

is calculated from the cross product of the two vectors $\vec{r}_{ji} - \vec{r}_{RP}$ for any two atoms j.

An improper rotation symmetry with $n = 2$ could represent either an $\mathbf{S_1}$ (reflection) or an $\mathbf{S_2}$ (inversion) symmetry. The distinction is made by computing the midpoints of the line segments joining each pair of symmetry-related atoms. If the midpoints all lie on the RP, to within a specified tolerance, then the symmetry is an inversion through the RP. However, if any midpoint does not lie on the RP, then the symmetry is a reflection symmetry. The corresponding plane of reflection is found by identifying two pairs of atoms related to each other by the reflection symmetry and drawing line segments connecting the members of these pairs. The plane is then defined by the midpoints of the two line segments and the RP. The same method is used to find the reflection plane of an improper rotation with $n > 2$; and in this case, the associated rotation axis is identified by drawing the two vectors from the RP to the two midpoints and taking their cross product.

Identification of Topological and 3D Symmetries in Multimolecular Complexes. Exchanging two chemically identical molecules that form part of a complex is considered an acceptable topological symmetry operation for eliminating repeated conformations generated by a conformational search algorithm and for identifying chemically equivalent atoms in algorithms for molecular construction. Thus, each global and local topological symmetry of a molecule belonging to a complex represents a local symmetry of the complex. For example, if a complex includes a benzene molecule, the rotational symmetry operations associated with the benzene do not change the conformation of the complex and hence represent local symmetry operations for the complex. This additional type of symmetry operation is handled by constructing the molecule name and atom sequences of a complex as the concatenation of the molecule names and corresponding sequences of its constituent molecules. Exchanging two chemically equivalent molecules in the complex then corresponds simply to interchanging their atom sequences within the name of the complex.

The detection of 3D symmetries for a complex proceeds as follows. First, all global and local 3D symmetries are detected for each molecule by the procedures described in previous subsections. Then the molecules that form the complex are grouped by their type, based upon their molecule names, and each group is assigned a unique code. For example, if a complex contains two benzene molecules and two cyclophanes, the benzenes might both be named "A" and the cyclophanes named "B". The individual molecules are then formed into a pseudomolecule by computing the center of coordinates of each molecule and connecting each pair of centers of mass with a pseudobond. Thus, the benzene/cyclophane complex just described would form a pseudomolecule A_2B_2 held together by 6 pseudobonds. Global symmetries are now identified in the pseudomolecule by the same method used at the molecular level, except that now pseudobond lengths are included in the conformation names. When a symmetry is detected at the level of the pseudomolecule, it is necessary to check whether it holds at the molecule level also. This is done by applying the pseudomolecule's putative symmetry operations and then checking whether the conformations of the overlapped molecules match. For example, a complex with 4 molecules of A and one of B might be arranged in space as follows:



The pseudomolecule possesses, among other potential symmetries, a possible 4-fold rotation with the axis through B and normal to the plane defined the complex. This potential symmetry is tested by applying the rotation operation 3 times and, for each step, determining whether any combination of atom−atom interchanges permitted by the local topological symmetries reduces the overall root-mean-square distance of corresponding atoms in each molecule below a user-defined threshold. If so, then the putative symmetry does exist.

When the pseudomolecule is linear, this procedure will not detect symmetries that involve rotation around the pseudomolecular axis or reflection through a plane containing this axis. Additional tests are therefore done to check for such symmetries. First, if each constituent molecule has a global reflection symmetry, and if their reflection planes are coplanar, then the complex possesses a global reflection symmetry through this plane. Similarly, if each constituent molecule has a global rotational symmetry such that all the rotations lie along the pseudobond axis, and if the symmetry orders are integer multiples of the lowest-order symmetry, then the complex possesses a global rotation symmetry of the lowest order.

Implementation and Evaluation. The algorithm was programmed in Fortran 77, except for the resonance detection and atom-charging codes, which include Python scripts. We find that the run time is roughly proportional to $O(m*n^2)$, where m is the number of starting points and n is the number of symmetry points. The method was tested for single molecules and multimolecular complexes with a variety of local and global symmetries, as now described.

RESULTS

This section presents illustrative results, focusing primarily on 3D symmetries. However, an illustration of the use of a topological symmetry for filtering out redundant conformations also is included. Note that every 3D symmetry necessarily corresponds to a topological symmetry.

Detection of 3D Symmetries. Application of the algorithm to the compounds and complexes shown in Figure 16 yields the symmetries reported in Table 1, as now reviewed.

Benzene in its energy-minimum conformation possesses three 2-fold rotations around axes joining *para* carbons and three 2-fold rotations around axes that bisect opposite bonds, for a total of 6 $\mathbf{C_2}$ symmetries; and each of these rotations can also represent a reflection through a plane perpendicular to the plane of the ring and containing the corresponding rotation axis, for another 6 $\mathbf{S_1}$ symmetries. There are also a global inversion symmetry ($\mathbf{S}_2$) and a global 6-fold rotation ($\mathbf{C}_6$) around a central axis perpendicular to the plane of the ring. The local symmetries correspond to the 6 symmetrically branching carbons encountered from the hydrogen starting atoms.

In the conformation shown, the **cyclic urea**[20] possesses a global 2-fold rotation symmetry whose axis is parallel to the

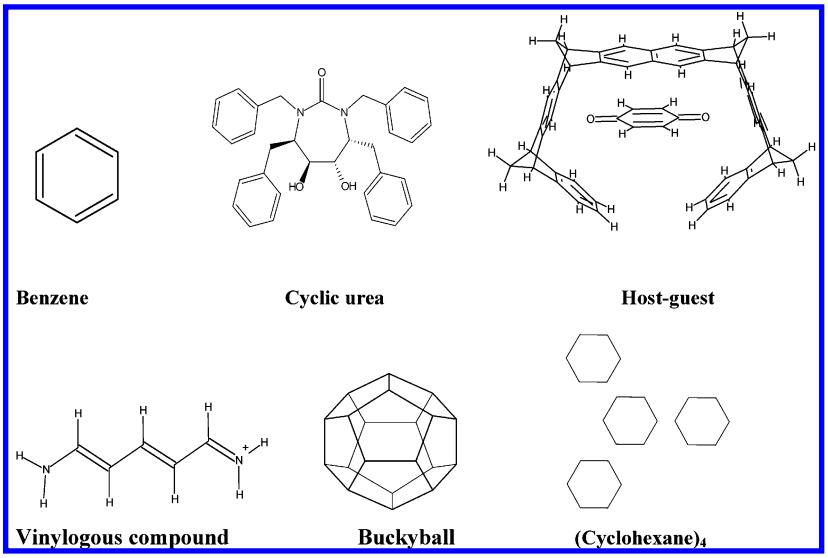


Figure 16. Molecules and complexes used to illustrate symmetry detection. Some hydrogens are not shown for clarity. Symmetries are listed in Table 1.

Table 1. Computed 3D Symmetries of Systems in Figure 16

molecule/ complex	with all hydrogens		without hydrogens	
	global symmetries	local symmetries	global symmetries	local symmetries
benzene	$6S_1$, C_6, $6C_2$, S_2	$6C_2$	$6S_1$, C_6, $6C_2$, S_2	$6C_2$
cyclic urea	$1C_2$	$8S_1$, $5C_2$	$1C_2$	$4S_1$, $5C_2$
$C_{20}H_{20}$	$15S_1$, $15C_2$, $6S_{10}$, $10S_6$, S_2	$60S_1$, $20C_3$	$15S_1$, $15C_2$, $6S_{10}$, $10S_6$, S_2	$60S_1$, $20C_3$
host−guest	$2S_1$, $1C_2$	receptor globals $2S_1$, $1C_2$ guest globals $2S_1$, $3C_2$, S_2 receptor locals $4S_1$ guest locals $2C_2$	$2S_1$, $1C_2$	receptor globals $2S_1$, $1C_2$ guest globals $2S_1$, $3C_2$, S_2, receptor locals none guest locals $2C_2$
vinylogous compd	$1S_1$, $1C_2$	$3S_1$	$1S_1$, $1C_2$	none
$(cyclohexane)_4$	$3S_1$, C_3	cyclohexane globals $4(3S_1, 3C_2, S_6, S_2)$ cyclohexane locals $4(6S_1)$	$3S_1$, C_3	cyclohexane globals $4(3S_1, 3C_2, S_6, S_2)$ cyclohexane locals none

carbonyl bond. Each benzene ring also contributes a 2-fold local rotational symmetry (C_2) with its axis parallel to the bond joining the ring to the rest of the molecule, and each of these local rotations can also be interpreted as a local reflection (S_1) symmetry. In addition, the 4 methylene groups each contribute a local reflection symmetry corresponding to the interchange of the two hydrogens, as illustrated in Figure 12. The final local symmetry is a 2-fold rotation whose associated symmetric branch point is the urea carbon. This local symmetry actually matches the global 2-fold symmetry, except that it does not involve the carbonyl oxygen. In some applications, it is convenient to merge such near-redundant local symmetries into their corresponding global symmetries, but the local symmetry is included here for completeness.

For the $C_{20}H_{20}$ **buckyball**, every pair of antipodal C−C bonds is associated with a global reflection symmetry with the plane bisecting the bonds. There are 30 such bonds and hence 15 S_1 symmetries. Each such pair of bonds is also associated with a C_2 rotation symmetry around an axis bisecting both bonds. Each pair of antipodal pentagonal faces is associated with an 10-fold rotation/reflection symmetry with the axis passing through the centers of the opposite pentagons, for a total of 6 S_{10} symmetries. (The C_5 symmetry associated with each pentagonal face is implicit in the corresponding S_{10} symmetry and hence is not listed separately.) Similarly, each pair of antipodal carbons is associated with a global S_6 symmetry, with implicit C_3 symmetries which are not listed in the table, for a total of 10 S_6 symmetries. Each atom is also related to its antipodal counterpart by a global inversion symmetry (S_2). The local symmetries are accounted for as follows. Every carbon is associated with 3 local reflection symmetries, where the carbon is the symmetric branch point and the plane of symmetry bisects a C−C−C bond angle. Each carbon also is the symmetric branch point for a 3-fold local rotational symmetry.

In the conformation shown, the **host−guest complex** as a whole possesses two global reflection symmetries. One plane splits the complex into left and right halves from the point of view in the figure, and the other splits it into front and back halves. There is also a global 2-fold rotation symmetry whose axis is normal to the plane of the guest and passes through its center. The symmetries of the receptor, which here are regarded as local symmetries of the complex, match those of the complex. However, the guest has additional symmetries. The planes of its two reflection symmetries are normal to the ring; one passes through the oxygens and the other bisects the ring's double bonds. The axes of its three 2-fold rotations run through the two oxygens, bisect the ring double bonds, and run normal to the ring and through its center. The other local symmetries of the complex correspond to 4 local reflections (S_1) bisecting the H−C−H moieties that project outward from the rest of the receptor and two local C_2 rotations of the guest, where the carbonyl carbons are the symmetric branch points. The **vinylogous**

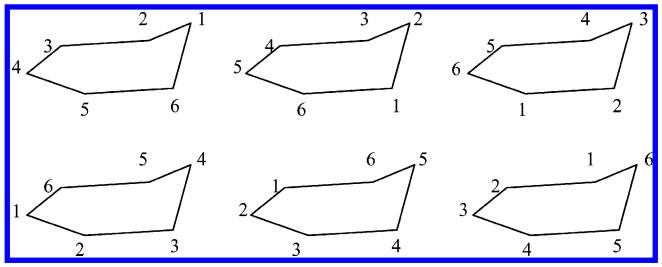


Figure 17. Six conformations of cyclohexane in the half-chair conformation, oriented uniformly to show their redundancy. Atom indices i_{file} are shown.

compound is included to highlight the importance of accounting for alternate resonance forms. When the alternate resonance form of this planar molecule is considered, it becomes clear that it possesses a global 2-fold rotation whose axis is colinear with the center C−H bond and a global 2-fold reflection whose plane contains the rotation axis and is normal to the plane of the paper. The compound also has 3 local reflection symmetries associated with the symmetric branching of the central carbon and the interchange of the hydrogens attached to the nitrogens.

The symmetric array of four cyclohexane molecules in the chair conformation, **(cyclohexane)$_4$**, possesses three global reflection symmetries whose planes each bisect one of the outer rings and the central ring as well as a global 3-fold rotation whose axis passes through the center of the central ring and is normal to the plane defined by the outer rings. Each cyclohexane molecule also possesses global and local symmetries which are regarded here as local symmetries of the complex. Thus, each cyclohexane molecule has three reflection symmetries whose planes contains the S_6 axis shown in Figure 13 and cut through two carbons separated by 3 bonds and three 2-fold proper rotations such as that shown in Figure 12. There are also a global inversion symmetry and the S_6 improper rotation shown in Figure 13. Each cyclohexane also possesses six local reflection symmetries corresponding to the exchange of the two hydrogens bonded to each carbon.

Acceleration by Omitting Hydrogens. The CPU times required to obtain the above results on a 2000+ Athlon computer running RedHat Linux range between 0.2 and 10 s each, depending upon the complexity of the system. We investigated the possibility of speeding the calculations by neglecting hydrogen atoms from the structures, since if only hydrogens are omitted, the missing hydrogens remain implicit in the rest of the molecule. The results are summarized in the right-hand side of Table 1, which shows that all global symmetries are still detected, and the only local symmetries that disappear are reflection symmetries associated with hydrogens bonded to a single parent atom. A substantial speedup is realized: the CPU times fall to 0.1−0.5 s for these cases, due to the reduction in the number of starting atoms to be examined and also the reduction in the number of branched atoms that need to be processed.

Conformational Filtering. As noted in the Introduction, topological symmetries are particularly useful as an aid to removing redundant conformations of a single molecule after application of a conformational search algorithm. Figure 17 illustrates this application with the simple example of 6 half-chair conformations of cyclohexane, each with a different out-of-plane carbon. Because all the carbons are chemically (topologically) equivalent, all 6 conformations should be regarded as equivalent. However, naively superposing the structures according to the atom indices (i_{file} in Methods) will yield large atom−atom distances, so the 6 conformations will be considered distinct. Detection of symmetries with the present method yields a 6-fold topological symmetry which equates the first atom sequence (shown in bold) with all the following sequences of the carbon atoms:

1	**2**	**3**	**4**	**5**	**6**
1	2	3	4	5	6
2	3	4	5	6	1
3	4	5	6	1	2
4	5	6	1	2	3
5	6	1	2	3	4
6	1	2	3	4	5

Here carbons are considered to be numbered consecutively around the ring, hydrogens are omitted for simplicity, and the identity is included as the first matching sequence. These equivalences can be used to correctly filter out redundant conformations by superposing each pair of conformations using all seven possible sets of atom−atom matchings and rejecting a trial conformation if the smallest atom−atom distances relative to another structure is less than a user-defined cutoff. By this means, 5 of the 6 redundant conformations in Figure 17 can be eliminated.

DISCUSSION AND CONCLUSIONS

The present algorithm identifies both topological and 3D symmetries and includes significant advances relative to previous methods. First, the speed is increased by reducing the number of potential symmetries that need to be checked in detail through the use of highly informative atom names based upon element, extended connectivity, and atomic charge. Additional time is saved by reducing the number of molecular traversals, in part by judicious selection of starting points, and it has been shown that further acceleration can be achieved, with minimal cost in information, by omitting hydrogens.

Importantly, the present method also provides functionalities beyond those of prior methods. Thus, it detects not only global symmetries but also local symmetries of interest in conformational analysis and molecular construction; it identifies symmetries for multimolecular complexes and is therefore well, and perhaps uniquely, suited to applications in supramolecular and host−guest chemistry; and it correctly identifies symmetries that become apparent when alternate resonance forms are considered. As a consequence, the method has a wide range of applications, including filtering out redundant conformations of host, guests, and bimolecular complexes during conformational analysis and free energy calculations; speeding conformational analysis and docking by reducing the sampling ranges of rotatable bonds that rotate locally symmetric groups; and automating the calculation of symmetry numbers for thermochemical applications.

ACKNOWLEDGMENT

The authors thank Drs. Robert Jorissen for helpful discussions, Hillary S. R. Gilson for extensive testing of the algorithm, Michael J. Potter for software support, and Chia-en Chang for providing useful test cases. This publication

was made possible by Grant Number GM62050 and Grant Number GM61300 from the National Institutes of Health. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the National Institutes of Health.

Note Added in Proof: The following paper published after submission of this work describes a version of Ivanov & Schüürmann's algorithm which is efficient for compounds with high topological symmetries. Ivanov, J. Molecular Symmetry Perception. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 596.

REFERENCES AND NOTES

(1) Wei, J. Molecular symmetry, rotational entropy and elevated melting points. *Ind. Eng. Chem. Res.* **1999**, *38*, 5019−5027.
(2) Brown, R. J. C.; Brown, R. F. C. Melting Point and molecular symmetry. *J. Chem. Educ.* **2000**, *77*, 724−731.
(3) Head, M. S.; Given, J. A.; Gilson, M. K. Mining Minima: Direct computation of conformational free energy. *J. Phys. Chem. A* **1997**, *101*, 1609−1618.
(4) Chang, C. E.; Potter, M. J.; Gilson, M. K. Calculation of molecular configuration integrals. *J. Phys. Chem. B* **2003**, *107*, 1048−1055.
(5) Kolossvary, I. Evaluation of the molecular configuration integral in all degrees of freedom for the direct calculation of conformational free energies: Prediction of the anomeric free energy of monosaccharides. *J. Phys. Chem. A* **1997**, *101*, 9900−9905.
(6) Ivanov, J.; Schüürmann, G. Simple algorithms for determining the molecular symmetry. *J. Chem. Inf. Comput. Sci*. **1997**, *39*, 728−737.
(7) Yip, R. K. K. A Hough transform technique for the detection of reflectional symmetry and skew-symmetry. *Pattern Recognit. Lett.* **2000**, *21*, 117−130.
(8) Yuen, K. S. Y.; Chan, W. W. Two methods for detecting symmetries. *Pattern Recognit. Lett*. **1994**, *15*, 279−286.
(9) Sun, C. Fast reflectional symmetry detection using orientation histograms. *Pattern Recognit. Lett.* **1999**, *16*, 987−996.
(10) Golender, V. E.; Drboglav, V. V.; Rosenblit, A. B. Graph potentials method and its application for chemical information processing. *J. Chem. Inf. Comput. Sci*. **1981**, *21*, 196−204.
(11) Faulon, J. L. Isomorphism, automorphism partitioning and canonical labeling can be solved in polynomial time for molecular graphs. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 432−444.
(12) Fraysseix, H. D. www.ehess.fr/centres/cams/papers/173.ps.gz.
(13) Cordella, L. P.; Foggia, P.; Sansone, C.; Vento, M. http://amalfi.dis.unina.it/graph/db/papers/vf-algorithm.pdf.
(14) Manning, J. Geometric symmetry in graphs, Ph.D. Thesis, Purdue University, 1990.
(15) Masinter, L. M.; Sridharan, N. S.; Carhart, R. E.; Smith, D. H. Applications of artificial intelligence for chemical inference. 13. Labeling of objects having symmetry. *J. Am. Chem. Soc*. **1974**, *96*, 7714−7723.
(16) Morgan, H. L. Generation of a unique machine description for chemical structures-a technique developed at Chemical Abstracts Service. *J. Chem. Doc*. **1965**, *5*, 107−113.
(17) Wipke, W. T.; Dyott, T. M. Stereochemically unique naming algorithm. *J. Am. Chem. Soc*. **1974**, *96*, 4834−4842.
(18) Gilson, M. K.; Gilson, H. S. R.; Potter, M. J. Fast assignment of accurate partial atomic charges: An electronegativity equalization method that accounts for alternate resonance forms. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1982−1997.
(19) http://www.mdli.com/downloads/public/ctfile/ctfile.jsp.
(20) Lam, P. Y.; Jadhav, P. K.; Eyermann, C. J.; Hodge, C. N.; Ru, Y.; Bacheler, L. T.; Meek, J. L.; Otto, M. J.; Rayner, M. M.; Wong, Y. N.; Chang, C.-H.; Weber, P. C.; Jackson, D. A.; Sharpe, T. R.; Erickson-Viitanen, S. *Science* **1994**, *263*, 380−384.

CI049966A