Simple Algorithms for Determining the Molecular Symmetry

Julian Ivanov*,†,‡ and Gerrit Schüürmann†

Department of Chemical Ecotoxicology, UFZ Centre for Environmental Research, Permoserstrasse 15, D-04318 Leipzig, Germany, and Bourgas Technological University, 8010 Bourgas, Bulgaria

Received January 28, 1999

Simple and efficient algorithms are presented for detecting the full set of molecular symmetry operations (rotation and reflection), for identifying the respective symmetry elements, and for assigning the molecular point groups. All molecular symmetry point groups are available. The algorithms can be easily generalized for any three-dimensional structure that can be defined by a set of vertexes $V(\nu_i)$, edges (pairs of connected vertexes) $E(\nu_i\nu_j)$, and the Cartesian coordinates of the vertexes.

1. INTRODUCTION

Molecular symmetry belongs to the fundamental properties of chemical structures. The type and degree of symmetry impose important restrictions on the quantum chemical wave functions describing the electronic structure of the molecular system, and as a consequence affect a number of physicochemical compound properties. A well-known example is given by the dipole moment, which is a vector property, such that its direction must be preserved when applying any symmetry operation of the molecular system. Hence, if a molecule has a proper rotation axis as part of its symmetry, the dipole moment must lie along this axis. If there is a plane of symmetry, the dipole moment must lie in this plane. A structure with an improper axis of symmetry (i.e., an axis of rotation−reflection, s.b.) cannot possess a dipole moment, because any improper symmetry operation reverses the direction of the vector. It shows how qualitative information about the dipole moment of a chemical compound can be obtained by knowledge of its molecular symmetry.

Another application of symmetry is optical activity. A molecule is optically active if it rotates the plane of polarized light passing through it. A criterion often used to determine if the molecule is optically active is to see if the structure is superimposable on its mirror image. Consideration of symmetry properties leads to a simple criterion to identify the optical activity of chemical compounds. If a molecular structure has an improper symmetry axis, it cannot be optically active, and vice versa, if it has no such an axis it is optically active.

It is thus of great importance for various branches of chemistry to have tools that enable us to find out all symmetry elements and to assign the symmetry point group for arbitrary types of chemical compounds. Several algorithms[1−5] have been proposed to detect the symmetry point group of the molecule or to discover different symmetry elements[6−8] in point set polygons, and polyhedra. Most of the algorithms for detecting the symmetry elements proceed by calculating the moments of inertia ($I_{x,y,z}$) or some other structural characteristics, followed by a reorientation of the molecule under investigation according to this specific frame of reference. After this coordinate transformation, the structure is checked for the presence of various (but finite number) symmetry elements, but to our best knowledge all of the presently published algorithms are confined to testing a finite number of symmetry operations. Hence, these algorithms do not guarantee the detection of the complete set of symmetry operations. In principle there is always a chance for the existence of a proper or improper axis of symmetry of order $\mathbf{n}+\mathbf{m}$, where $\mathbf{n}$ is the maximum number of checked symmetry operations and $\mathbf{m} \geq 1$, with $\mathbf{n}$ and $\mathbf{m}$ being integer numbers.

As a consequence, so-called optimal algorithms[9] have been proposed for detecting the full set of symmetry operations for a number of geometric entities. It has been shown that, for polygons and polyhedra with connected, planar surface graphs, the complexity of such algorithms depends linearly on the number of vertexes, $\mathbf{n}$, while for point sets and general polyhedra the corresponding dependence can be expressed as the function: $\Theta[n \log(n)]$. However, the authors do not provide an explicit and complete solution for finding the full set of symmetry operations for the 3D objects.

A nice approach[10] has been developed recently for detecting the set of rotational symmetry operations of polyhedral objects. It has explored the isomorphism of the graphs and incorporated some geometrical constraints to estimate the rotational symmetry, similar to the strategy in the present paper. However, the authors do not consider the reflection symmetry.

In the present work we propose simple algorithms for detecting the full set of rotation and reflection molecular symmetry operations, and for identifying the according symmetry elements and finally assigning the molecular point group for molecular structures of arbitrary kind. Moreover, the new set of algorithms can be easily generalized for other three-dimensional (3D) objects, provided that these can be defined by a set of vertexes $V(\nu_i)$ and edges (pairs of connected vertexes) $E(\nu_i\nu_j)$, together with the Cartesian coordinates of the vertexes.

2. DEFINITIONS

In this section we briefly recall some major elements from group theory together with the notation commonly applied, which is applied in the remainder of the paper.

† UFZ Centre for Environmental Research.
‡ Bourgas Technological University.

10.1021/ci990322q CCC: $18.00
Published on Web 05/18/1999

A symmetry operation is a transformation of the representation of an object such that the transformed representation is indistinguishable from the initial one with respect to any physically observable property, which implies that in particular the distances between all pairs of points of the object are preserved.

A symmetry element is a geometrical entity (point, line, or plane) with respect to which one or more symmetry operations may be carried out. Since symmetry operations can be defined only with respect to symmetry elements, and at the same time the presence of symmetry elements can be demonstrated only by applying that appropriate symmetry operations exist, it is necessary to consider the related types of elements and operations together.

Four types of symmetry elements are considered for characterizing molecular symmetry (axis of symmetry, plane of symmetry, center of symmetry, and rotation−reflection axis of symmetry).

We say that a body has an axis of symmetry (also proper axis of symmetry) if rotation about this axis by 360/*n* degrees gives a physically indistinguishable configuration from the original position, where **n** is called the order of the axis (**n** is an integer). The symbol for an **n**-fold rotation axis is $\mathbf{C_n}$. A proper axis of order **n** corresponds to **n** symmetry operations: C_n, C_n^2, ..., C_n^n (=E symbol for operation identity).

A given structure has a symmetry plane if the operation reflection of all atoms through that plane gives a configuration physically indistinguishable from the original one. The symbol for a symmetry plane is σ. All atoms lying in the plane constitute special cases, since the operation of reflection through the plane does not affect their position at all. Hence, planar molecules have at least one plane of symmetry (molecular plane). As an important consequence, all atoms outside a symmetry plane must occur in even numbers, since each atom on the one side of the plane must have a twin on the other side. The presence of a symmetry plane generates only one symmetry operation, in contrast to the **n**-fold axis of rotational symmetry.

A molecule has a center of symmetry (or center of inversion) if the operation of inverting all the atoms through the center gives a configuration indistinguishable from the original one. The symbol for the center of symmetry is ***i***. Applying the operation of inversion does not move an atom lying in the center of symmetry. All other atoms must occur in pairs, since each must have a twin with which it is exchanged upon inversion. The center of symmetry corresponds to one symmetry operation.

The last type of symmetry element is a rotation−reflection axis (or improper axis) of symmetry, symbolized by $\mathbf{S_n}$. The presence of an improper axis $\mathbf{S_n}$ implies that rotation by 360/*n* degrees about the axis, followed by reflection of all the atoms in a plane perpendicular to the axis, gives a configuration physically indistinguishable from the original one. $\mathbf{S_n}$ with even **n** corresponds to *n* symmetry operations and always implies the presence of a $\mathbf{C_{n/2}}$ axis. When **n** is odd, $\mathbf{S_n}$ corresponds to 2**n* symmetry operations.

The full set of symmetry operations of a given isolated molecule forms a mathematical group, called a point group.

3. THE CODING ALGORITHM

The problem of identifying and characterizing molecular symmetry is closely related to testing the equivalence of different ways of labeling sequences (paths) of atoms in graph theory. The different assignments of the vertexes of the chemical graph are used to describe the symmetry. A good reference on this topic is.[11] The authors define: "Two numberings of a graph are equivalent if the connection table derived from one can be made identical with that derived from the other by rearrangements of the rows and of the connection lists within each row." And also: "The symmetry group of a graph is the set of all permutations whose corresponding numberings yield ordered connection tables identical with that of the reference numbering." If properties other than the connection table are used, other types of symmetry groups may be defined. This is the basic idea we use to discover all molecular symmetry operations for a given molecular structure.

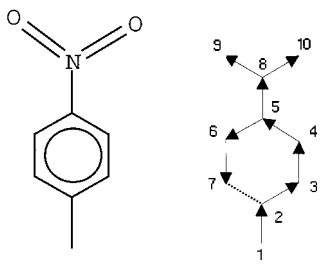

Figure 1. Structure formula with the associated oriented spanning tree of *p*-nitrotoluene.

Since we work with molecular structure representations (connection tables, for example) instead of real chemical structures, it is necessary to have a tool that generates correct representations. A structure representation is correct if it conveys the constitution of the chemical compound, i.e., of its atom types and bond types as well as of the connectivity of the structure. Two representations are equivalent if they are correct representations of the same compound. The difference between two equivalent representations results from different numberings of the nodes in the chemical graph. Since there are **n!** different ways to assign the labels, there are **n!** correct connection tables for a given structure (where **n** is the number of nodes in the chemical graph). Hence, it is necessary to have a set of rules that control the order in which the structure is described. Once this order has been determined, the description itself is a relatively simple process.

In the present work, the ordering of the labels is obtained on the basis of the oriented spanning tree of the chemical graph, constructed by the Depth First Search (DFS) procedure.[12] The process of numeration begins with an initial ordering of the vertexes of the chemical graph according to their connectivity. The spanning tree (Figure 1) starts at a vertex with the lowest connectivity. Then one goes to an adjacent vertex with lowest connectivity, then to a third vertex adjacent to the second one, and so on. Vertexes are visited only once. When all adjacent nodes have been visited, one returns to the last vertex on the walk which still has neighbors that have not yet been visited, and continues by taking the neighbor with lowest connectivity. Edges connecting the current vertex with a visited one that is not its ancestor are termed ring closure edges. Thus, the nodes of the chemical graph are ordered with respect to the sequence in which they are visited. When all vertexes of the structure are visited, a linear code is generated. The name (code) of

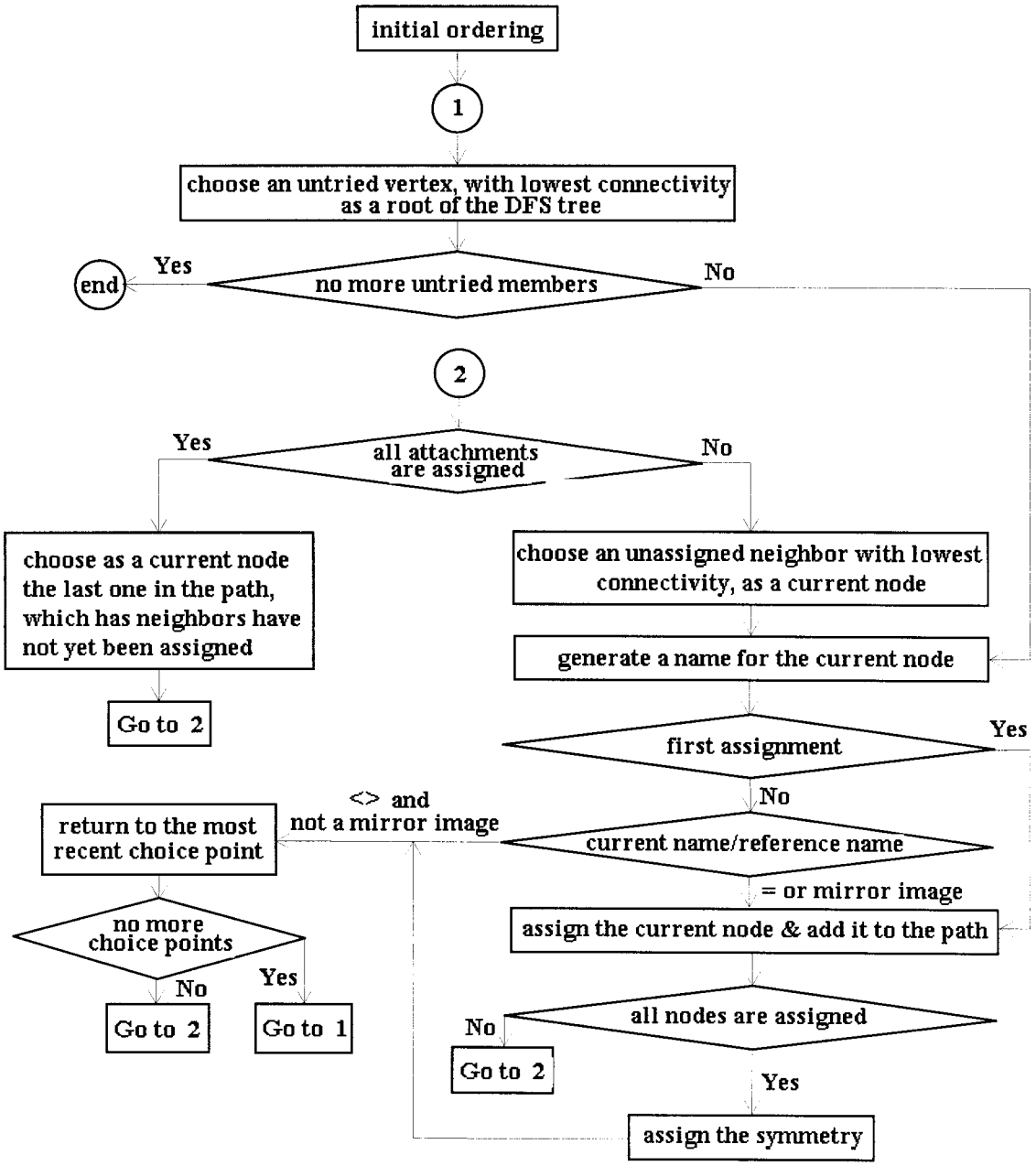

Figure 2. Flow chart of the coding algorithm.

the chemical graph is formed by the sequence of atom labels according to the ordering they have in the DFS tree. Except for the first atom, each atom label is followed by the bond type and DFS tree number of the ancestor (ancestors, in case of ring-closure edges). The symbols for bond types which are used in the present work are as follows: −, single bond; =, double bond; #, triple bond; •, aromatic bond. The root of the tree has no ancestor, so the first element of the name is the symbol of the atom type of the first node ("C" for the structure in Figure 1) followed by the atom type of the second vertex "C", the symbol for bond type between the second atom and its ancestor, "−", the DFS number of the ancestor, "1", and so on. Thus, for example, the complete topological code of the structure in Figure 1 is C/C−1/C•2/C•3/C•4/C•5/C•2•6/N−5/O=8/O=8. Here, the symbol "/" is used just for clarity to delimit the description of the particular atoms. The hydrogen atoms are ignored only to make the examples shorter. The first generated linear name is used as a reference name. To limit the combinatorial size of the problem, comparing at any step the current name with the reference, one controls the process of coding. If a difference is found, the procedure is put back to the previous step. In the case of a choice between **n** atoms with one and the same connectivity, **n** names are generated. When an identical name with the reference one is generated due to a choice point, then the nodes interchanged by this choice are defined to be equivalent (with respect to the criteria on which the name is based). The main steps of the algorithm are illustrated in the flowchart in Figure 2. Several versions of this naming algorithm are described in the literature[10,13−15].

As mentioned above, there are **n!** possible numberings of a given compound. For any of these **n!** numberings, a linear name can be juxtaposed. In the case of an asymmetric structure, there are **n!** different names. If the given structure possesses equivalent nodes, any different name corresponds to a set of different equivalent numberings of the nodes of

the chemical graph. All **n!** assignments of the vertexes can be present as follows:

$$\begin{pmatrix} \text{name \#1 } \{P_{1\,1}, P_{1\,2}, ..., P_{1\,m}\} \\ \text{name \#2 } \{P_{2\,1}, P_{2\,2}, ..., P_{2\,m}\} \\ \hline \text{name \#}N \{P_{N\,1}, P_{N\,2}, ..., P_{N\,m}\} \end{pmatrix}$$

where $\{P_{i\,1}, ..., P_{i\,m}\}$ denotes the different permutations of the nodes, corresponding to the $\mathbf{i}^{th}$ name, and N indicates the number of different names. In other words, any of these permutations corresponds to a symmetry operation (with respect to the criteria on which the name is based). Hence, finding all molecular symmetry operations is equivalent to determining all possible permutations of the vertexes of the chemical graph according to a given linear name (the reference one). The main difference from refs 13−15 is the aim for which the algorithm is used, while the goal of refs 13−15 is to generate a unique linear name, i.e., to take the "best" name from the set of N possible different names (see the scheme above). The purpose of the present work is to find out all possible equivalent assignments of the nodes of the given structure, which correspond to one linear name. Here, we do not need the "best" name, so for the sake of brevity we just use the first generated (reference) name. In other words, we must determine the first row $\{P_{1\,1}, P_{1\,2}, ..., P_{1m}\}$ of the scheme above. Actually, in refs 13−15 the symmetry property of structures is used to decrease the combinatorial task, needing at least $\mathbf{N + m - 1}$ (not **n!**) steps, while the nature of the task in the present paper requires only **m** steps of the algorithm for determining the molecular symmetry operations.

4. THE THREE-DIMENSIONAL (3D) NAME

Since we consider the different numberings corresponding to a given name, the criteria on which the name is based are crucial for the algorithm to characterize molecular symmetry. Restriction to atom types, bond types, and connectivity as described above leads to a topological code of the chemical structure, thus confining symmetry information to molecular topology. Incorporation of the parity of stereocenters results in the derivation of the according stereo name of the chemical structure.[13] To account for the 3D nature of the compounds, it is necessary to include in the linear name some 3D structural characteristics such as dihedral angles,[13,15] bond angles, bond distances, etc. in view of the fact that the dihedral angles and the stereoconfigurations of the chemical structures are incorporated in the 3D name not only to increase the accuracy but mainly to account for the reflection symmetry nature of the chemicals (see below). These structural characteristics are considered as a particularly important part of the 3D code. For the sake of simplicity of the description below, we do not treat the 3D part referring to bond distances and bond angles explicitly. This corresponds to accepting that the structures under consideration are not distorted; i.e., the bond distance and bond angle parts of the 3D name are the same in any case.

The so-called conformational name of the chemical structure is formed by the sequence of values of all possible dihedral angles in the spanning tree (without the dihedral angles, which include the ring-closure bonds). In the present work, a dihedral angle is defined by four sequentially connected atoms. The value of the angles varies between π and $-\pi$. A special treatment exists when the value is on the boundary (π, $-\pi$). Any rotations by 180° about the axis passing the central bond will change the sign of the dihedral angle, so in this case the latter is taken together with its absolute value. The torsion angles are ordered following the sequence of the edges in the DFS tree. When there are several dihedral angles about one and the same central edge, these torsion angles are ordered with respect to the lowest DFS numbering of the nodes.

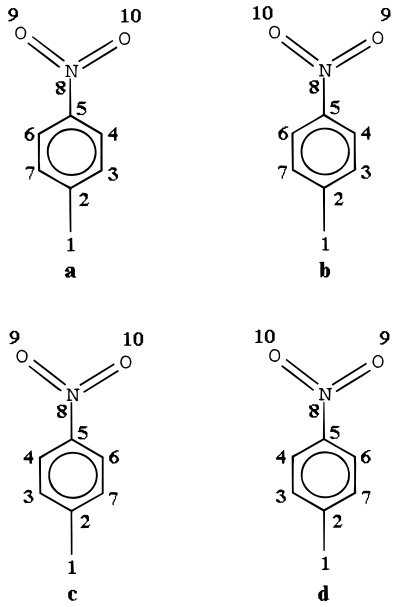

Figure 3. Four equivalent numberings of the atoms of the topological name of *p*-nitrotoluene.

Table 1. The Four Conformational Names (Columns **a**−**d**) Associated with Four Numberings of the Atoms of *p*-Nitrotoluene, Which Correspond to Four Topological Symmetry Operations

	conformational names			
dihedral angles	**a**	**b**	**c**	**d**
1-2-3-4	180	180	180	180
2-3-4-5	0	0	0	0
3-4-5-6	0	0	0	0
3-4-5-8	180	180	180	180
4-5-6-7	0	0	0	0
8-5-6-7	180	180	180	180
4-5-8-9	180	0	0	180
4-5-8-10	0	180	180	0
6-5-8-9	0	180	180	0
6-5-8-10	180	0	0	180

An example with *p*-nitrotoluene is shown in Figure 3, where four different but equivalent topological numberings are presented corresponding to one topological name, with associated dihedral angle information (Table 1) corresponding to more specific conformational names, which are specified by the columns of the table. The existence of four different topological numberings means that there are four topological symmetry operations. Inspection of the graphs shows further that the relation between the conformational names (columns **a**−**d** in Table 1) is $\mathbf{a = d \neq b = c}$. There are two different numberings corresponding to the reference conformational name (column **a** in Table 1); hence, there are two 3D symmetry operations. The first symmetry operation is identity E, and the second one is the rotation by 180° around the axis passing through nodes 1 and 8 (see

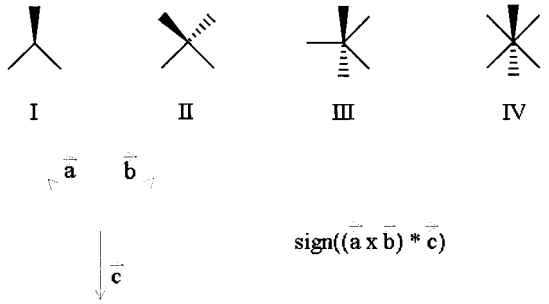

Figure 4. Four structures (**I**−**IV**) with atoms attached to a central atom, where the longest linear chain of atoms consists of only three atoms.

Figure 3). The topological symmetry operation **a** → **b** corresponds to the internal rotation of the NO_2 group by 180° about bond 5−8; hence, it cannot be considered as a 3D symmetry operation.

Figure 4 shows structures with atoms attached to a central atom, where the longest linear chain of atoms has only three atoms. In such cases, the 3D name (called configuration name) is formed as a sequence of the values of the stereoconfigurations for all possible ordered subsets of three atoms attached to the central atom, whose total number is given by $\binom{n}{3}$ with **n** being the number of attachments. The stereoconfiguration is defined as the sign of the cross-product of the three vectors, (**a** × **b**)***c** (see Figure 4). Here, the possible values are −1 (cross-product is negative), +1 (cross-product is positive), or 0. The last result is obtained when the three vectors lay on a plane or when the vector product (**a** × **b**) is undefined, in which case the two vectors are lying on a line.

Compound **I** in Figure 4 has six equivalent topological numberings (corresponding to the reference topological name). Here, the total number of ordered subsets of three attached atoms is $\binom{3}{3} = 1$ (only 1, 2, 3); hence, the according 3D code has only one element (the stereoconfiguration of the three attachments). Three of the numberings have configuration name +1, and the other three, −1 (Figure 5A). This means that there are three 3D-symmetry operations, each of which is given as a rotation by 120° around the axis passing the central node and being perpendicular to the plane on which are lying the three attachments.

For structure **II** in Figure 4, the configuration name has four elements (the ordered subsets of three attached atoms are 1,2,3/1,2,4/1,3,4/2,3,4). The number of equivalent topological assignments of the nodes for this structure is 24. Twelve of them have a reference configuration name −1/1/−1/1, and the other 12 1/−1/1/−1 (Figure 5B). The according symmetry elements are four C_3 axes passing the four edges and three C_2 axes bisecting the valence angles of the structure. For the last two structures **III** and **IV**, the according topological numberings are 120 and 720, respectively, and the number of equivalent numberings corresponding to the reference configuration name are 6 for **III** and 24 for **IV**. At the end of this paragraph, we must mention that for structures with two and three atoms neither dihedral angle nor stereoconfiguration can be defined. The number of equivalent topological assignments of the vertexes is quite sufficient (along with the bond distance part of the name for three atoms molecules) for determining the molecular symmetry operations. In both cases, it cannot be more then 2.

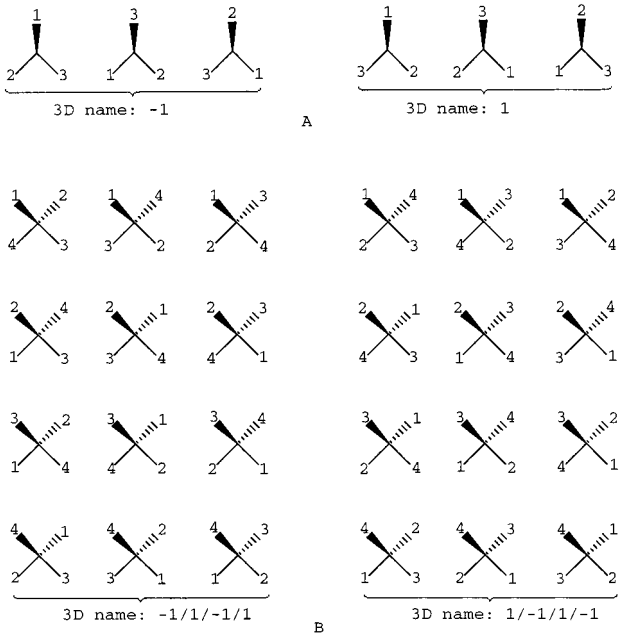

Figure 5. Equivalent numberings of the atoms, representing all 3D-symmetry operations, of (A) structure **I** and (B) structure **II**.

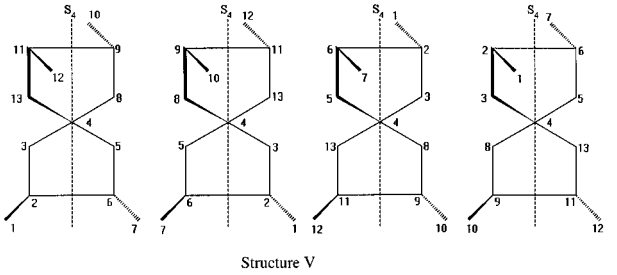

Figure 6. Four equivalent numberings of the atoms of structure **V** corresponding to 3D symmetry operations that belong to the reference topological name.

As the dihedral angles and stereo configurations convey the rigidity of the 3D structures, all of the above examples of symmetry operations are related to rotational symmetry only. In this context, all symmetry elements we can detect (with respect to the introduced criteria for 3D conformational names) are the proper rotation symmetry axes. To account for the reflection symmetry, we have to consider the remaining three operations of molecular symmetry inversion (***i***), reflection through a plane (σ), and improper rotation ($\mathbf{S_n}$). As σ is $\mathbf{S_1}$ and ***i*** is $\mathbf{S_2}$, for simplicity any of these operations can be regarded as an improper rotation.

Consideration of reflection symmetry does not require inclusion of new structural characteristics in the 3D conformational name. Since any improper operation contains a reflection through a plane, it just has to be checked whether the current 3D name is the mirror image of the reference one. The mirror image of the conformational/configurational name is defined as the name in which any of the elements differs from the corresponding one in the reference name only with respect to sign, except for the boundary values 0, π (where $0 = -0$ and $\pi = -\pi$).

This is illustrated by the structure given in Figure 6, which has eight numberings corresponding to one and the same topological name. Four of them conform to internal rotations of the structure; hence, they cannot be regarded as molecular symmetry operations. The other four assignments (presented

Table 2. The Four Conformational Names (Columns **a**−**d**) Associated with the Four Numberings of the Atoms of Structure **V**, Which Corresponds to Four 3D Symmetry Operations

	conformational names			
dihedral angles	**a**	**b**	**c**	**d**
1-2-3-4	−120	−120	120	120
2-3-4-5	0	0	0	0
2-3-4-8	−120	−120	120	120
3-4-5-6	0	0	0	0
8-4-5-6	120	120	−120	−120
4-5-6-7	−120	−120	120	120
3-4-8-9	−120	−120	120	120
5-4-8-9	120	120	−120	−120
4-8-9-10	120	120	−120	−120
4-8-9-11	0	0	0	0
8-9-11-12	120	120	−120	−120
8-9-11-13	0	0	0	0
10-9-11-12	120	120	−120	−120
10-9-11-13	−120	−120	120	120

in Figure 6) represent four 3D symmetry operations. The first two conformational names (columns **a** and **b** in Table 2) are identical; hence, the according symmetry operations are proper rotations about the axis (denoted as S_4 in Figure 6) by 180°. The last two 3D names (columns **c** and **d** in Table 2) are the mirror image of the reference name (column **a**); hence they correspond to improper symmetry operations. As a matter of fact, rotation about the S_4 axis by 90° or −90° and reflection through the plane perpendicular to this axis and passing through the center of the mass (atom #4) will leave the body of the structure on a position physically indistinguishable from the reference position. The molecular symmetry point group of the structure in Figure 6 is S_4, because the only symmetry element is the S_4 axis. As is well-known from group theory, the symmetry element S_4 also contains C_2.

For the structures in Figure 5 (examples for the configuration name), the numberings of the nodes in the left three columns correspond to one and the same name for the appropriate structures; i.e., they correspond to proper symmetry operations. The numberings in the right three columns have a configuration name that is a mirror image of the reference one (the left name); hence, they correspond to improper symmetry operations with respect to the numberings in the left three columns. Thus, for structure A in Figure 5, any of the numberings in the right part can be obtained from the left numberings as a reflection through a plane which bisects the valence angle between two of the attachments and contains the third one. There are three such planes σ_v in addition to the proper C_3 symmetry axis, and the appropriate symmetry point group is C_{3v} (NH_3, CH_3X, etc. belong to this point group).

The assignments in the right part of structure B in Figure 5 can be obtained from the left numberings by reflection through planes which are defined by the central atom and any two of the attachments. There are six such σ planes. Besides them, there are three C_2 and four C_3 proper axes as mentioned above. Any of the four C_2 axes is collinear with a S_4 axis. Examples for this symmetry point group (T_d) are CH_4, tetrahedron, adamantane skeleton, etc.

In some cases, the identical names are also mirror images. Thus, for the structure in Figure 3, names **a** and **d** are identical but they are also mirror images, while any $\mathbf{a_i} = -\mathbf{d_i}$ (all the values are on the boundaries 0 or π). It means that in this case the proper and improper symmetry operations are equivalent: after carrying out the two symmetry operations separately, the two resulting numberings will be identical. In fact, both rotating the reference structure about the axis passing atoms #1 and #8 (Figure 3) by 180° and reflecting through the plane passing atoms #1 and #8 and perpendicular to the molecular plane yield one and the same numberings (name **d** in Table 1).

5. IDENTIFICATION OF THE MOLECULAR SYMMETRY ELEMENTS AND THEIR COORDINATES

In the previous two sections, the naming algorithm was described, and criteria for the 3D conformational name were introduced that allow finding all (rotation and reflection) molecular symmetry operations. However, the according symmetry elements are not yet completely determined. The only thing we can say so far is, that a certain symmetry element is responsible for proper or improper symmetry operation. The order of the symmetry axes and their coordinates in the 3D space are still unknown. The aim of the present section is to determine these geometrical entities.

Note first that the structure under study is checked for linearity. So the following considerations apply only to nonlinear compounds.

At any step of the naming algorithm we identify a symmetry operation, i.e., a new numbering equivalent to the reference one. Starting from this mapping of the nodes of the chemical graph and their Cartesian coordinates, application of simple rules of vector analysis enables the identification of the symmetry elements in a quite straightforward way.

As is well-known from group theory, any symmetry element passes the center of mass (CM). Hence, the coordinates of the first point of any symmetry element are perfectly determined by eqs 1:

$$X_{CM} = \frac{\sum_{i=1}^{n} x_i}{n};\ Y_{CM} = \frac{\sum_{i=1}^{n} y_i}{n};\ Z_{CM} = \frac{\sum_{i=1}^{n} z_i}{n} \qquad (1)$$

5.1 Determining the Proper Rotation Symmetry Elements. We have a proper rotation symmetry operation if and only if the current 3D conformational name is identical to the reference one. The according symmetry element can only be an axis. Hence, it is necessary to determine the order of the axis and an associated collinear nonzero vector. To this end, we follow the next scheme:

Choose an atom i which does not lie on CM and compare Current Number of i [CN(i)] to Reference Number of i [RN(i)]. Now there are two cases.

I. RN(i) = CN(i). It means that the chosen atom does not change its position upon application of the symmetry operation; i.e., it lies on the searching axis. Since atom i does not lie on CM, the collinear vector ($\vec{CV}$) is [CN(i) − CM]. Next the order of the axis has to be determined. The procedure is repeated again by choosing an atom for which case II (see below) is true.

II. RN(i) ≠ CN(i): the chosen atom changes its position when the symmetry operation is carried out. Without restriction of generality, we can state that RN(i) is transformed into CN(i)=N_1, and it follows that RN(N_1) ⇒ CN(N_1)=N_2, RN(N_2) ⇒ CN(N_2)=N_3, and so on along this path,

until $CN(N_n) = RN(i)$ is reached. The length (**n**) of this path is the order of the axis. If the collinear vector ($\vec{CV}$) is not yet determined, it can be easily found by eq 2, were ***n*** is the order of the axis.

$$\vec{CV} = \sum_{i=1}^{n} \overline{[CN(N_i) - CM]} \qquad (2)$$

If $|\vec{CV}| = 0$, all vectors from the sum of eq 2 lie on a plane or on a line ($\boldsymbol{n} = 2$). For that specific situation, two possible cases are regarded: $\boldsymbol{n} > 2$ and $\boldsymbol{n} = 2$. In the first case, the coordinates of the vector are calculated from the vector product of the first two vectors of the sum, applying eq 3:

$$\vec{CV} = \overline{[CN(N_1) - CM]} \times \overline{[CN(N_2) - CM]} \qquad (3)$$

If $\boldsymbol{n} = 2$, the information we have is not sufficient to determine $\vec{CV}$, since the two atoms and CM lie on line **L**, perpendicular to the searching axis, and there is an infinite number of such axes. In this case, the vector $\vec{C}_1 = \overline{[CN(N_1) - CM]}$ is calculated (collinear with **L**), and the entire procedure is repeated again by choosing an atom which does not lie on **L**. If the second run leads again to a situation where the vector cannot be defined directly, $\vec{CV}$ can then be calculated as the vector product of the vectors $\vec{C}_1$ and $\vec{C}_2$. The latter is calculated exactly like $\vec{C}_1$ in the second run of the procedure.

5.2 Determining the Improper Symmetry Elements. The condition for the existence of an improper symmetry operation is that the current conformational name must be a mirror image of the reference name. Since the center of inversion coincides with the center of mass (CM), the only task left is to identify the symmetry operation inversion. The plane of symmetry, σ, is defined by two vectors starting at CM that do not lie on one line. For the definition of the last symmetry element, rotation−reflection improper axis, it is necessary to find a vector collinear to it, and to specify the order of the axis. The calculation of the coordinates of two vectors that define the reflection plane perpendicular to the axis is a trivial task.

The steps of the algorithm are as follows:

Choose an atom *i* which does not lie on CM.

I. If $RN(i) = CN(i)$, the chosen atom lies on the reflection plane. If the plane is not yet defined, calculate a new vector. When the second vector is calculated, check the angle between the two vectors. In case this angle is more then 0° and less than 180°, the plane is perfectly defined; otherwise, go to the start of the procedure. If an improper symmetry axis exists, there must be found at least one pair of suitable noncollinear vectors defining the reflection plane.

II. If $RN(i) \neq CN(i)$, count the order ***n*** of the improper element, following the path in a way similar to the algorithm for the proper rotation axis.

II.1 If $\boldsymbol{n} = 2$, there are only two possibilities:

• **The current symmetry operation is inversion.** Since each of the atoms has a twin (except for an atom coinciding with CM), the middle point of any pair of twins coincides with the CM of the structure, provided that the symmetry operation inversion is applied.

• **The current symmetry operation is a reflection through some plane** σ**.** If for a pair of atoms its middle point differs from CM, the current symmetry operation is a reflection through some plane. Consequently, such a middle point determines a vector, which can be used for the definition of the reflection plane.

II.2 If $\boldsymbol{n} > 2$, the current symmetry element can only be $\mathbf{S_n}$.

In this case for any sequence pair of atoms of the path $CN(N_i)$, $CN(N_{i+1})$, the middle point lies in the reflection plane and in particular is different from CM. Together with CM, such a middle point defines a vector. Any pair of such vectors, which do not lie in the same line, can be used to define the reflection plane. The vector product of these two vectors determines the $\mathbf{S_n}$ axis.

Since any $\mathbf{C_n}$ axis generates **n** symmetry operations, and any $\mathbf{S_n}$ axis corresponds to **n** or **2*n** operations for **n** being even and odd, respectively, the order of the according axis is the maximal value of **n**, calculated from the different symmetry operations corresponding to this axis.

6. ASSIGNING THE MOLECULAR POINT GROUP

As we determine the full set of symmetry operations (the naming algorithm) and completely define the according symmetry elements (the types and their 3D coordinates), we have generated all information that is needed to assign the molecular point group of a given structure. The only task left is to check the relationships between the different symmetry elements.

At any step of the coding algorithm, we determine a new symmetry operation and the corresponding element. The latter is compared with each member of the set of already discovered symmetry elements. Now, the properties of the different point groups are considered.

6.1 The Infinite Point Groups (R3, $\mathbf{C_{\infty v}}$**,** $\mathbf{D_{\infty h}}$**).** If the structure is a single atom, the point group is **R3**. If the structure is linear, the possible point groups are $\mathbf{C_{\infty v}}$, $\mathbf{D_{\infty h}}$. Here, a sufficient criterion is the number of equivalent numberings of the nodes of the chemical graph, corresponding to the reference name, which is one for $\mathbf{C_{\infty v}}$, and two for $\mathbf{D_{\infty h}}$.

6.2 The Cubic Point Groups (T, $\mathbf{T_d}$**,** $\mathbf{T_h}$**, O,** $\mathbf{O_h}$**, I,** $\mathbf{I_h}$**).** If at some step of the algorithm two different axes with order $\mathbf{n} > 2$ are discovered, the structure under investigation belongs to one of the cubic point groups.

If at least $\mathbf{C_3}$, $\mathbf{C_4}$, or two different $\mathbf{C_4}$ axes are discovered, the according point group is **O** or $\mathbf{O_h}$. The presence of a center of symmetry ***i*** determines $\mathbf{O_h}$, and the absence of ***i*** determines **O**.

Similarly, if there are $\mathbf{C_3}$, $\mathbf{C_5}$, or two different $\mathbf{C_5}$ axes, the point group is $\mathbf{I_h}$ in case ***i*** is also present; otherwise, the structure belongs to symmetry group **I**.

The symmetry point group is **T**, $\mathbf{T_d}$, or $\mathbf{T_h}$, if at the end of the procedure the highest order of the different discovered symmetry axes is 3. Then ***i*** defines the $\mathbf{T_h}$ point group. The absence of ***i*** determines $\mathbf{T_d}$ or **T**; here, the existence of plane of symmetry determines $\mathbf{T_d}$; otherwise, **T** is present.

6.3 The Axial Point Groups ($\mathbf{C_n}$**,** $\mathbf{C_{nv}}$**,** $\mathbf{C_{nh}}$**,** $\mathbf{D_n}$**,** $\mathbf{D_{nd}}$**,** $\mathbf{D_{nh}}$**,** $\mathbf{S_n}$**).** These groups are characterized by a highest order ($\mathbf{n} \geq 2$) axis named the principle symmetry axis. To recognize them, one must take into account the presence of different $\mathbf{C_2}$ axes, a reflection plane, and its position relative to the principal axis.

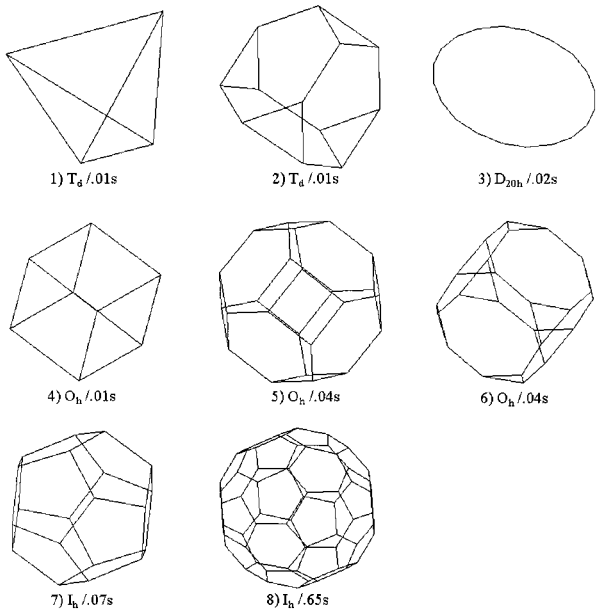

Figure 7. Eight geometric 3D structures together with the associated symmetry point groups and the CPU times needed for the symmetry calculations.

If the structure under study has only one symmetry axis (the principle one), the possible point groups are $\mathbf{C_n}$, $\mathbf{C_{nv}}$, $\mathbf{C_{nh}}$, $\mathbf{S_n}$. When there is no plane of symmetry or improper rotation, the point group is $\mathbf{C_n}$. If there is a reflection plane, we consider two cases: (i) the plane is perpendicular (named σ_h) to the principal axis, $\mathbf{C_{nh}}$; and (ii) the plane (named σ_v) contains the principal axis, $\mathbf{C_{nv}}$. In the second case, there should be **n** such symmetry planes, and the angle between any two neighboring planes is 360/**n**.

When the only symmetry element is an improper rotation axis, the point group is $\mathbf{S_n}$ ($\mathbf{S_2} \infty \mathbf{C}_i$). Note that with $\mathbf{S_n}$ point groups the number of improper rotations is even **n**; otherwise, the structure belongs to $\mathbf{C_{nh}}$ ($\mathbf{S_n} \infty \mathbf{C_{nh}}$, when **n** is odd).

If besides the principal axis, there are other $\mathbf{C_2}$ symmetry axes, the given structure belongs to one of the so-called dihedral point groups: $\mathbf{D_n}$, $\mathbf{D_{nd}}$, $\mathbf{D_{nh}}$.

The absence of any reflection plane defines a $\mathbf{D_n}$ point group. If there is a symmetry plane σ_h, the point group is assigned as $\mathbf{D_{nh}}$. This point group also possesses **n** σ_v planes. Any of these planes contains the principal axis and one of the $\mathbf{C_2}$ axes. The angle between any pairs of neighboring planes σ_v or $\mathbf{C_2}$ axes is 360/**n**.

When there are reflection planes but there is no an σ_h, the point group is $\mathbf{D_{nd}}$. There should be **n** such reflection planes, called σ_d, and each of these planes contains the principal axis and bisects the angle between two neighboring $\mathbf{C_2}$ axes.

Finally, there are another two special point groups $\mathbf{C_1}$ and $\mathbf{C_s}$, which are characterized only by the trivial symmetry element identity **E**. For these groups, the 3D conformational name corresponds to only one numbering of the atoms. If the conformational name is also its own mirror image, implying a planar structure, the point group is $\mathbf{C_s}$; otherwise, it is $\mathbf{C_1}$.

7. EXAMPLES AND DISCUSSION

Applications of the above-described algorithms are developed in C for UNIX workstations (IBM, Silicon Graphics), and in Borland Pascal and C++ for IBM-compatible PCs. The computer applications are incorporated as tools in the OASIS system for computer-aided structure−property relationship investigation,[16,17] and in the ChemProp system

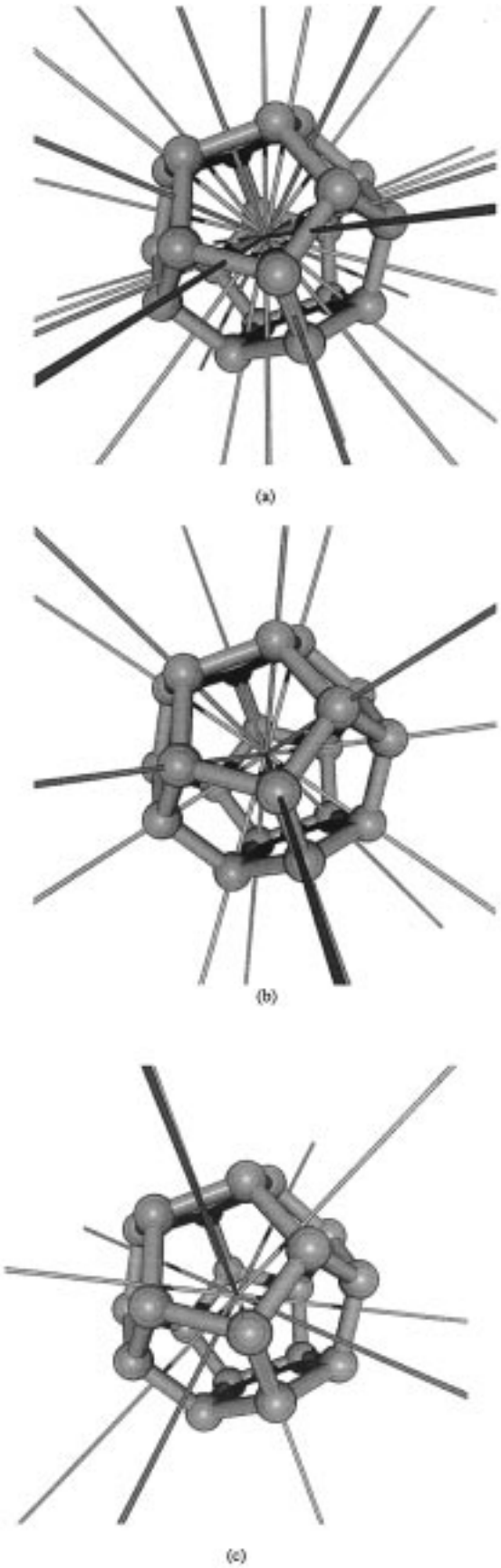

Figure 8. Symmetry axes of dodecahedron ($\mathbf{I_h}$): (a) $15C_2$ axes; (b) $10S_6$ axes; (c) $6S_{10}$ axes.

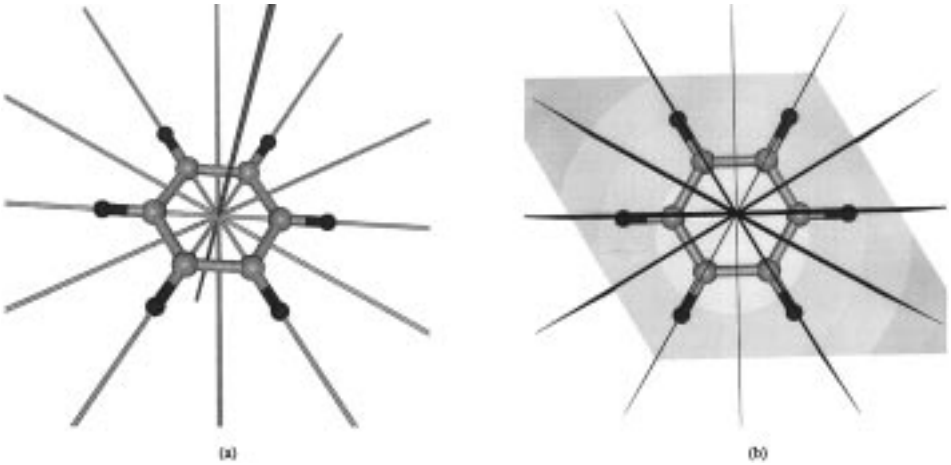

Figure 9. Symmetry elements of benzene ($\mathbf{D}_{6h}$): (a) $6C_2$ and C_6 axes; (b) $6\sigma_v$ and σ_h reflection planes.

for calculating physicochemical compound properties directly from chemical structure.[18] Although the current algorithm for identifying symmetry operations (the coding algorithm) shows a quadratic dependence of computation time on molecular size, the CPU times are quite attractive for chemical structures with small to moderate numbers of atoms and edges (up to 300). In Figure 7, some examples of geometric structures with high but finite molecular symmetry are shown together with the (calculated) point group and the associated CPU time needed on the Silicon Graphics workstation OCTANE.

It is important to note that our algorithm for determining the molecular or geometric symmetry does not need any specific orientation of the structures under study, and that (to our best knowledge) there are also no other limitations with recognizing any symmetry point group as outlined above. This contrasts with other symmetry algorithms implemented in current software packages. An example is given by Dmol[19] and Gaussian 94,[20] where (according to the manuals) a proper identification of rotational groups requires Cartesian *z* to be the principal axis, and where the horizontal symmetry plane must be the Cartesian *xy* plane for recognizing C_s, C_{nh}, and D_{nh}.

The current symmetry module implemented in MOPAC 2000[5] and its handling of cubic groups give another example. Cubic systems are oriented by identifying atoms of the set nearest to the center of symmetry. If there are 4, 6, 8, 12, or 20 of these, and the number of equidistant nearest neighbors is 3, 4, 3, 5, or 3, respectively, then the atoms are at the vertexes of one of the Platonic solids (tetrahedron, octahedron, cube, icosahedron, dodecahedron). Therefore, all atoms of the set lie on high-symmetry axes. The first atom is selected and used to define the *z*-axis.

Structures 7 (pentagonaldodecahedron) and 8 (buckmisterfulleren) represent more demanding examples. To find out all 120 symmetry operations (E, $12C_5$, $12C_5^2$, $20C_3$, $15C_2$, *i*, $12S_{10}$, $12S_{10}^3$, $20S_6$, 15σ), to calculate the 3D coordinates (and order of the axes) of the according symmetry elements, 31 axes ($15C_2$, $10S_6$, $6S_{10}$), and 15 σ planes, and finally to assign the point group ($\mathbf{I_h}$), the symmetry calculation needs 0.07 and 0.65 s CPU time, respectively. Although this is certainly acceptable, there should still be room for improvement, because we have not yet put any efforts in optimizing the code to increase the calculation speed. Figure 8a−c shows all symmetry axes for the pentagonaldodecahedron as calculated by the computer application. The pictures are taken directly from the interactive graphics tools of the ChemProp system.[18] As the reflection planes of the pentagonaldodecahedron are difficult to see, a further example with reflection symmetry planes is given in Figure 9 for benzene ($\mathbf{D_{6h}}$ point group).

The algorithm for determining the molecular symmetry operations at any step compares the different geometrical entities (bond distances, valence, and dihedral angles) of the structure under study. These two procedures can be used to address further aspects of molecular geometries such as the breaking of symmetry associated with minor geometrical distortions, which is an object of further research.

8. CONCLUSIONS

To our best knowledge, the procedures described above represent the first set of algorithms that allow, without any restriction, the automatic recognition of all symmetry point groups associated with geometric entities. The examples analyzed so far show acceptable calculation times with the code implemented in computerized form, which will allow us to handle large lists of medium-sized molecules within quite moderate CPU times. However, the calculation time increases in quadratic order with molecular size, and future investigations may concentrate on optimizing the code to make it feasible also for larger molecules. Another route of future work may be the application for quantum chemical calculations with the additional task to adapt the wave functions to the identified molecular symmetry.

REFERENCES AND NOTES

(1) Calvert, J. B. *Am. J. Physiol.* **1963**, *31*, 659.
(2) Hollas, J. M. *Symmetry in Molecules*; William Clowes & Sons Ltd.: London, 1972.
(3) Cotton, F. A. *Chemical Applications of Group Theory*, 2nd ed.; Wiley: New York, 1981.
(4) Levine, I. N. *Quantum Chemistry*; Prentice Hall: Englewood Cliffs, NJ, 1991; pp 322−341.
(5) MOPAC 2000, J. J. P. Stewart; Fujitsu Limited: Tokyo, Japan, 1999.
(6) Yuen, K. S. Y.; Chan, W. W. *Pattern Recognit. Lett.* **1994**, *15*, 279−286.
(7) Yip, R. K. K.; Lam, W. C. Y.; Tam, P. K. S.; Leung, D. N. K. *Pattern Recognit. Lett.* **1994**, *15*, 919−928.
(8) Sun, C. *Opt. Eng.* **1997**, *36*(4), 1073−1077.
(9) Wolter, J. D.; Woo, T. C.; Volz, R. A. *Vis. Comput.* **1985**, *1*, 37−48.

(10) Jiang, X. Y.; Bunke, H. *CVGIP: Graphical Models Image Processing* **1992**, *54*(1), 91−95.
(11) Masinter, L. M.; Sridharan, N. S.; Carhart, R. E.; Smith, D. H. *J. Am. Chem. Soc.* **1974**, *96*(25), 7714−7723.
(12) Tarjan, R. E. *SIAM J. Comput.* **1972**, *1*, 146.
(13) Wipke, W. T.; Dyott, T. M. *J. Am. Chem. Soc.* **1974**, *96*, 4834.
(14) Ivanov, J.; Karabunarliev, S.; Mekenyan, O. *J. Chem. Inf. Comput. Sci.* **1994**, *34*(2), 234−243.
(15) Karabunarliev, S.; Ivanov, J.; Mekenyan, O. *Comput. Chem.* **1994**, *18*(2), 189−193.
(16) Mekenyan, O.; Karabunarliev, S.; Bonchev, D. *Comput. Chem.* **1990**, *14*, 193.
(17) Mekenyan, O.; Karabunarliev, S.; Ivanov, J.; Dimitrov, D. *Comput. Chem.* **1994**, *18*(2), 173−187.
(18) Schüürmann, G.; Kühne, R.; Kleint, F.; Ebert, R.-U.; Rothenbacher, C.; Herth, P. In *Quantitative Structure−Activity Relationships in Environmental Sciences−VII* (Chen, F., Schüürmann, G., Eds.) SETAC Press: Pensacola, FL, **1997**; pp 93−114.
(19) DMol Version 960, Biosym Technologies: San Diego, CA, 1996.
(20) Gaussian 94, Programmer's Reference, Carnegie Office Park, Building Six, Pittsburgh, PA 15106, 1995.

CI990322Q