Shape-Similarity Measures for Molecular Bodies: A Three-Dimensional Topological Approach to Quantitative Shape-Activity Relations

Paul G. Mezey

Mathematical Chemistry Research Unit, Departments of Chemistry and Mathematics, University of Saskatchewan, Saskatoon, Canada S7N 0W0

Received June 19, 1992

Precise shape characterization is essential in the study of correlations between the shapes of formal molecular bodies and chemical and biochemical properties within the QShAR (quantitative shape-activity relations) approach to molecular engineering and pharmaceutical drug design. The notion that similarity of three-dimensional molecular bodies can be quantified and measured using algebraic topological methods and combinations of geometrical and topological techniques forms the basis of the earlier shape group approach. In another recent development, the density domain approach (DDA) to chemical bonding replaces the conventional graphs of structural bond diagrams with a more revealing, quantum chemically correct sequence of all topologically distinct families of density domains (bodies enclosed by isodensity contours). In the present study, the shape group method (SGM) and the associated numerical shape codes of molecules are reformulated within the framework of the density domain approach. The DDA framework also provides a rigorous topological definition of functional groups of organic chemistry. Based on the general GSTE similarity principle (geometrical similarity as topological equivalence) and the RBSM (resolution-based similarity measures) technique introduced earlier, a new family of numerical measures of both large-scale and small-scale shape similarity within sequences of computer-generated three-dimensional molecular models are proposed. Special topological shape analysis techniques are discussed, combining the T-hull method of generalized convexity analysis and simple molecular representations, such as the fused-sphere van der Waals surfaces designed to study both local and global features of molecules.

INTRODUCTION AND MOLECULAR TOPOLOGY BACKGROUND

In this section a brief, qualitative description of some of the concepts and terminology of molecular topology is given, with references to textbooks and the original works. The subjects discussed in this section form the basis of the new developments described in the following sections.

In conventional quantitative structure-activity relations (QSAR) analysis,[1-4] the concept of molecular structure is usually interpreted as the graph defined by the chemical bonds or on a more sophisticated level, as the three-dimensional, stereochemical bonding pattern of the molecule. Whereas the above approaches have been quite successful in many applications, the more recent advances in 3D molecular modeling, in particular, rigorous molecular shape analysis techniques applied to formal molecular bodies and formal molecular surfaces[5] have set the stage for new developments. Representations of formal molecular bodies and surfaces provide more detailed and more chemically relevant information than simple molecular graphs and stereochemical bond structures and give a more faithful description of actual molecular recognition and interaction processes where the peripheral regions of molecular bodies and not the formal bonds play a prominent role. This realization has led to the development of the topological shape group method (SGM) for rigorous shape description[6-10] and to the subsequent proposal of the quantitative *shape*-activity relations (QShAR) approach to molecular engineering and pharmaceutical drug design.[6,7,11-15] The QShAR methods focus on the study of correlations between the 3D shapes of formal molecular bodies and chemical and biochemical properties.

In QShAR studies, the main apparent difficulty is finding appropriate shape characterization. We assume that establishing the presence or lack of correlations between faithful numerical shape descriptors and various activity data is a trivial computational task. Consequently, in this paper the emphasis is on numerical shape characterization.

The QShAR approach has been formulated within a topological framework. We use the term "geometrical object" for entities which can be described adequately by geometrical means and the term "topological object" for entities requiring a topological description. Topological methods have important advantages over the more conventional, geometrical techniques of shape analysis. Clearly, molecules are not geometrical but topological objects,[11] since most small geometrical changes do not alter the chemical identity of molecules. In dynamic molecular processes, for example, in conformational rearrangements or chemical reactions, some geometrical shape features can change while some of the essential, topological shape properties remain invariant. These invariant shape features and the families of nuclear arrangements (the domains in configuration space) where these shape features are preserved can be characterized by algebraic topological means. The algebraic topological methods are suitable for algorithmic, nonvisual shape analysis by computer programs.

One of the most versatile tools for rigorous molecular shape analysis is the shape group method.[6-10] In its simplest form, it is applicable to the shape analysis of molecular contour surfaces. The molecular surface is partitioned into domains, defined, for example, by local curvature properties: locally convex (D_2), locally saddle type (D_1), and locally concave (D_0) domains. The pattern and mutual relations among these domains define several topological groups, the *shape groups of the molecule*. The shape groups are the two-dimensional, one-dimensional, and zero-dimensional homology groups of truncated surfaces, obtained from the original surface $G(a)$ by removing all domains of a given type (for example, D_2, D_1, or D_0). Homology groups are powerful tools of algebraic

0095-2338/92/1632-0650$03.00/0

topology, suitable to extract the essential shape properties of objects.[16] In particular, the ranks of the homology groups, called *Betti numbers*,[16] provide concise numerical characterization of the topology of the given object. In most chemical applications to date, the one-dimensional homology groups of the D_2-type truncation of molecular contour surfaces provided the most efficient molecular shape characterization. The shape groups (which are algebraic groups not determined by the point symmetry groups of the molecules) can be obtained algorithmically for various computer-generated molecular surfaces, such as isodensity contours or contours of molecular electrostatic potentials. Instead of curvature properties, the surface patterns can be defined by many other criteria, for example, one may use the interpenetration patterns of isosurfaces of two or more physical properties, such as electrostatic potential, local spin density, and the magnitude of electronic density gradient.[8]

The ranks of the shape groups (their Betti numbers) are important topological invariants, providing a numerical, nonvisual shape characterization. For example, taking an isodensity contour surface $G(a)$ for some density value a and the local curvature domains of this surface, the shape groups and their Betti numbers are invariant within some density intervals (a_i, a_{i+1}). The same applies if a generalized concept of convexity is applied[7,10] by comparing the local curvatures along the molecular surface to a curved test object, such as a sphere of radius $1/b$: the shape groups are invariant within ranges (b_j, b_{j+1}) of the reference curvature b. Since all these intervals are of positive, noninfinitesimal length, there are only finitely many intervals (a_i, a_{i+1}) and (b_j, b_{j+1}). Consequently, there are only finitely many topologically different shape groups for the entire charge density distribution of the molecule, taking all physically relevant density values a and reference curvatures b into account. By listing the relevant topological invariants (usually their one-dimensional Betti numbers) for all these finite number of shape groups, a list of numbers, that is, a *numerical shape code*, is obtained. This shape code provides a detailed and rigorous shape characterization of the entire electronic charge density of the molecule.[7] The important aspect is that one does not have to select a special density value a to represent the molecule, all physically relevant density values are considered, and a complete shape description is given.

A similar consideration applies when taking into account the conformational flexibility of molecules.[9] The shape groups are also invariant to small conformational changes, hence within any conformational domain in configuration space only a finite number of different shape groups occur. This property of shape groups can be exploited in dynamic shape characterization of molecules within conformational domains, leading to the dynamic shape space approach.[9]

The similarity of the shapes of three-dimensional molecular bodies can be quantified and measured by comparing their numerical shape codes. The series of one-dimensional Betti numbers can be ordered into a vector $\mathbf{a}$ (if only one parameter, a or b is considered) or into a matrix (if both density parameter a and reference curvature parameter b are considered). Subsequently, any of the standard methods of comparing matrices and vectors are applicable to obtain numerical measures for their similarity.[11-15] This task can be carried out automatically by the computer, without involving visual inspection of the shapes and the subjective elements of judging similarity by human observers. The algorithmic shape similarity evaluation by the computer is fully reproducible.

The shape groups are invariant to small variations in

(a) the electron density contour value a and the associated variations in the actual isodensity contours $G(a)$

(b) the reference curvature value b

(c) nuclear configuration K

This invariance is an example of a more general, underlying principle of topological shape analysis. The general GSTE similarity principle (geometrical similarity as topological equivalence) can be given in a simple form proposed earlier.[11]

In order to describe the above general concept, we shall use an example. Two contour surfaces which have curvature domain patterns that can be transformed into one another by some continuous transformation (called homeomorphism in topology) are topologically equivalent. It is not necessary that the two domain patterns agree geometrically. As long as such transformation exists, this implies some geometrical similarity. If the pattern of the locally convex and concave domains are considered as the essential shape properties, then the topology of this pattern extracts these essential shape features, while disregarding the "incidental", detailed geometrical properties. The geometrical similarity of the two contour surfaces is reflected in a topological equivalence within the framework of the given topology. This is not the only possible choice, and topologies can be constructed by many other methods, for example, by considering the pattern of ranges of electrostatic potential on fused spheres van der Waals surfaces.[17-19] A topological equivalence between such patterns indicates a different type of similarity between two molecules or between two conformations of the same molecule.

Topology provides a rather general approach toward the study of similarity simply by focusing on the essential. What is to be regarded essential may be chosen rather freely, and this choice can be used to define the actual topology considered. If the selected features of two objects are similar enough, then this similarity appears as a topological equivalence of the two objects, within the constraints of the given topology. The choice of topology reflects the initial decision as to which features will be considered essential. For example, geometrical curvature properties of contour surfaces can be used to define domains on these surfaces. Subsequently, the pattern of these domains is analyzed topologically. The similarities of the *geometrical* features and mutual arrangements of these domains are detected as a *topological equivalence*. The general strategy for topological shape analysis can be formulated as the *GSTE principle*: geometrical similarity is treated as topological equivalence.[11]

A special treatment applies if the various choices of topologies are interrelated. Let us assume that the topologies are comparable in the mathematical sense,[16] that is, a cruder-finer relation exists between any two of the members of the given family of topologies. (A set open in the cruder topology is also open in the finer topology.) Some of these topologies find the two objects similar (topologically equivalent) whereas some other (finer) topologies may distinguish the objects (by these topologies the two objects are not equivalent). Then a grade of similarity can be associated to pairs of objects, based on the position of the two subfamilies of topologies within the cruder-finer hierarchy. This approach is based on a formal "topological resolution", providing a measure of similarity. A special case of such resolution-based similarity measures (RBSM) has been described in detail in ref 20.

There is a general scheme for the application of the GSTE principle, common to all topological descriptions, using the concept of *(P,W)-similarity*.[11] When applying this concept, one has to decide on the context of similarity, that is, one has to choose and define a *shape representation* and a *shape*

descriptor. These definitions require the specification of the two main components of the actual (P,W)-similarity:

(a) Choice of the *shape representation*, P, taken as the physical or geometrical property or model P selected to represent molecular shape.

(b) Choice of a suitable topological tool, W, to be used as the actual *shape descriptor* of P.

The shape representation, property, or model, P, may be chosen in many different ways. For example, P may be taken as an electronic charge isodensity contour surface $G(a)$ for a specified density value a; as a whole family of such isodensity surfaces within some range (a_1, a_2) of density values; as a fused sphere van der Waals surface for a specified set of formal atomic radii or for a whole family of such fused-sphere surfaces for a range of radii; as a geometrical description of the backbone of a protein structure as a ribbon or as a space curve; or as a sequence of simple geometrical objects such as cylinders, sheets, and rectangular blocks, representing schematically the main building blocks of proteins.

There is a considerable freedom in the choice of the topological shape descriptor W as well. If the shape representation P is the pattern of various shape domains on an isodensity contour where the pattern is defined by locally convex, concave, and saddle-type domains *relative* to a sphere of a specified curvature b, then the shape descriptor W (topological tool W) can be chosen as the shape groups[6-10] (homology groups of the truncated contour surface, obtained by eliminating domains of specified curvature properties) or shape matrices[5,21] (defined by the neighbor relations, types, and relative sizes of the domains).

Two molecules M_1 and M_2 are (P,W)-similar if for the common shape representation

$$P_1 = P_2 \tag{1}$$

their shape descriptors are topologically equivalent:

$$W_1 \sim W_2 \tag{2}$$

that is, if a homeomorphism (a one to one, onto, continuous mapping with a continuous inverse) exists between W_1 and W_2. The (P,W)-similarity is an equivalence relation among molecular arrangements. The above two molecules are (P,W)-equivalent, which can be expressed as

$$M_1 \ (P,W) \ M_2 \tag{3}$$

The above general scheme allows one to construct algorithms for nonvisual, algebraic shape characterization. For a given (P,W) choice, a whole family of possible geometrical arrangements of different molecules may have a common actual realization of the shape descriptor W (e.g., a common shape group or a common shape matrix). All these arrangements can be collected into a family, and the entire family can be represented by the given, common realization of the shape descriptor W. Each realization of W is called a (P,W)-*shape type*, denoted by $\tau_{(P,W)}$, or simply by τ if the (P,W) pair is implied from context. With the exception of some degenerate cases, there are only a finite number of different shape types τ_i. Having a common shape type τ_i is a topological equivalence relation, representing a geometrical similarity of the chosen shape representations. Usually, the shape types τ_i are specified by algebraic methods (e.g., by a group or a matrix) in terms of numbers. Consequently, the (P,W) shape-similarity technique provides a nonvisual, algebraic, algorithmic shape description, suitable for automatic, computer characterization of shapes and for the numerical evaluation of 3D shape similarity.

DENSITY DOMAIN SHAPE GROUP METHOD AND CHEMICAL IDENTITY OF FUNCTIONAL GROUPS

The density domain approach (DDA) to chemical bonding has been proposed[5] as a tool that is able to describe the full, three-dimensional bonding pattern within molecular bodies. Whereas formal chemical bonds of a molecule, usually imagined as lines interconnecting the nuclei, provide only a molecular skeleton, actual molecules are held together by a three-dimensional, fuzzy body of electronic charge distributions. Since the actual chemical bonding is not constrained to a set of lines in space, a more appropriate description of chemical bonding must take into account the full molecular body. Electronic charge densities, regarded as 3D bodies, and their isodensity contour surfaces $G(a)$ offer an alternative to the classical stereochemical bond diagrams and a more realistic model for the representation of chemical bonding.

A maximum connected part of an isodensity contour surface and the volume enclosed by it is called a density domain, DD.[5] Formally, a density domain is a maximum connected component of a level set $F(a)$ of the electronic charge density for a given threshold value a. [A level set $F(a)$ contains all points where the density is greater than the threshold a.] The boundary of the level set $F(a)$ is a single, closed part of the isodensity surface $G(a)$.

The density domain approach is based on the shape variations of the $G(a)$ contour surface as a function of the density parameter a. At a high electronic density value, that is, for a sufficiently large value of the contour parameter a, the isodensity surface $G(a)$ is composed of several disconnected, nearly spherical surfaces. At high density values each separate component of $G(a)$ encloses one nucleus. As the contour parameter a decreases, the isodensity contours expand and various parts of the contour surface $G(a)$ become connected. The sequence according to which various parts of $G(a)$ become connected provides information on the 3D shape of the molecular electronic charge density, indicating the pattern by which the electron density bonds the molecular fragments together. Some molecular fragments retain a separate closed contour within a wide interval of density values a; usually, these fragments have well-established chemical identity as a "functional group". A gradual decrease of value a eventually leads to the interconnection of all parts of $G(a)$. The isodensity contour becomes a single envelope surface, surrounding all the nuclei of the molecule. For very small isodensity contour values a, the surface $G(a)$ becomes a nearly spherical balloon.

It is possible to follow the shape changes of the contour surface $G(a)$ as the parameter value a scans the interval $(0,\infty)$. The most essential changes are the connections between parts of $G(a)$; these changes of $G(a)$ are topologically significant according to the metric topology[16] of the 3D space (that is, according to the customary topological notions of connectedness in a space with the usual, Euclidean distance concept). These changes occur only at a small, finite number of selected contour values a_i. The shapes of various parts of $G(a)$ and the pattern of their connections, as a function of the parameter a, provide a new, systematic approach to 3D chemical bonding.

The density domain approach (DDA), proposed earlier,[5] provides a 3D topological tool for a more comprehensive description of chemical bonding. By varying the contour parameter a over the range $0 \leq a \leq \infty$, various parts of the isodensity contour $G(a)$ become connected or disconnected. The isodensity surface values a_i at which such connections or disconnections occur are characteristic to the given molecular configuration and the electronic state. Of course, if the isodensity surfaces are calculated by some ab initio technique, then these a_i values and the associated shapes of the isodensity

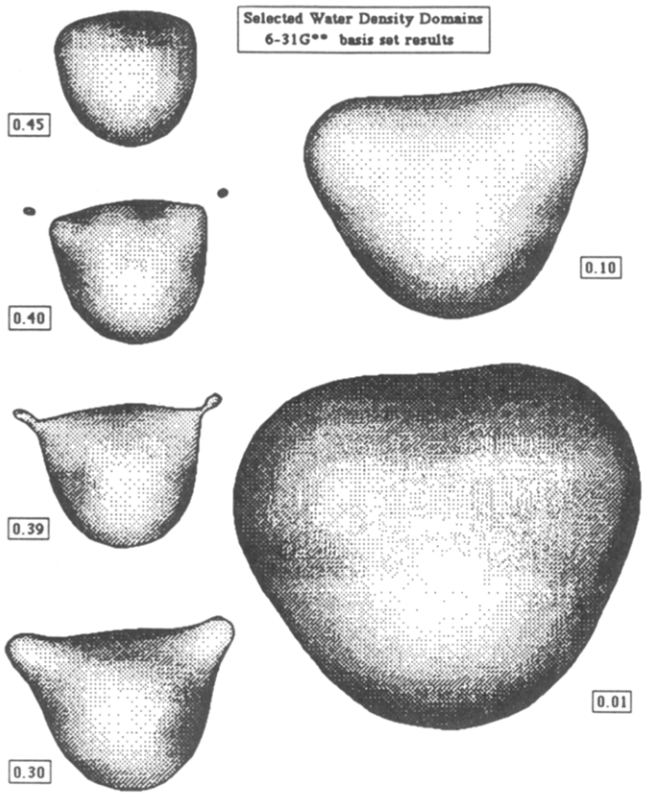

Figure 1. Selected density domains of the water molecule. Density values along isodensity contours are given in atomic units.

contours $G(a_i)$ are also dependent on the level of quantum chemical approximation applied. At a given value a, any connected part of the contour $G(a)$ represents a molecular fragment that can be regarded bound together by chemical bonding at electronic density level a. One may consider each connected part of the level set $F(a)$ as the representative of the body of a part of the molecule, chemically bonded at electronic density level a. A fragment that is connected at a higher density value a is held together by a stronger chemical bonding than one that is connected only at some lower threshold value a for $G(a)$. At very high density values a, one finds only disconnected atomic spheres; at such extreme density levels only individual atomic fragments can be regarded as undivided entities. By contrast, at low isodensity contour values a, all the nuclei of the entire molecule are enclosed by a single contour surface $G(a)$, hence at such electronic density level a the entire molecule can be regarded as chemically bonded, held together by a single envelope. The associated concept of chemical bonding is three-dimensional; it is based on the connection between the various parts of the molecular body at various density values. The actual bonding is not modeled by some infinitely thin "bonding channel", such as the conventional model for a formal chemical bond represented by a line. Instead, chemical bonding is attributed to a whole region of the space, occupied by electronic charge cloud, where the density is at or above some threshold value a. Note that the electronic density is not negative anywhere, (in contrast to the formal bonding and antibonding regions of Berlin diagrams[22] which have different signs), and the electronic charge present in any part of a molecular neighborhood can be regarded as bound to the molecule. The density domain approach describes the relative contributions of various regions of space to chemical bonding of molecular fragments, by considering the pattern of their stepwise interconnection as the positive isodensity contour parameter a is varied.

In Figure 1 selected isodensity contour surfaces of the water molecule are shown, as calculated using the GAUSSIAN 88 program[23] with a 6-31G** basis set, followed by the GSHAPE 90 molecular shape analysis program developed in our laboratory.[24] The shape analysis computations require only a small fraction of CPU time needed for the ab initio computation. For each contour density value a, the density domains DD(a) are the maximum connected components (separate pieces) of the volumes enclosed by the isodensity contour surface $G(a)$. According to the metric topology of the 3D space, there are only two topologically different families of density domains. The first of these is represented by the isodensity contour of density value $a = 0.45$ (all quantities are in atomic units). This surface has only one component, it is a topological sphere, and the only nucleus it encloses is that of the oxygen atom. At a lower density value, contours about the hydrogen nuclei appear, and a topologically different family of density domains is obtained, illustrated by the contour of density $a = 0.40$. This topological type of three disjoint topological spheres persists only in a rather narrow density range, and at density $a = 0.39$ one finds again an isodensity surface of a single connected component, a topological sphere. This sphere, as well as all other isodensity contours with contour density value less than 0.39, encloses all three nuclei of the water molecule. One may say that at and below the density value of 0.39 the water molecule is bound together, whereas at the density value of $a = 0.40$ the water molecule is disconnected. The essential features of the full, 3D bonding pattern of water molecule can be characterized by the density domains of a sequence of three topological objects, a sphere, a set of three spheres, and a sphere from which two, the first and the last, are topologically equivalent. The first and third topological objects are single DDs, whereas the second topological object is the combination of three DDs.

It is advantageous to combine the DD and the shape group methods. One approach is to carry out shape group analysis for each range of DDs separately, providing a natural classification of electron density intervals according to the 3D bonding pattern. A somewhat simpler approach is based on the selection of a representative density value from each density interval of DD topological invariance and carrying out a detailed shape group analysis only for the selected isodensity surfaces. For example, one may select the average density value within each density range. Clearly, the first approach is superior, since the curvature properties can change considerably within each such density interval. For example, in Figure 1 all the displayed water isodensity surfaces at and above $a = 0.39$ are nonconvex, but the relative nonconvexity is more prominent for the $a = 0.39$ surface than for the surface at $a = 0.01$. At an even lower a value, the corresponding isodensity surface is already convex, and at very low a values the surface becomes essentially spherical in the geometrical sense. The shape groups are suitable to detect much more subtle shape differences.

The choice of considering only the DDs or carrying out in addition a shape group analysis of the DDs leads to two levels of shape characterization and similarity evaluation.

On a simpler level, shape is characterized by the sequence of families of DDs as they occur while the a value is gradually decreased. Two molecules with the same sequence are similar according to the DD topology. This approach to similarity can be refined and quantified by taking the ratio of the number of mismatches along the sequences and the length of the longer sequence. A finer measure is obtained if the matches along the sequences are weighted by ratio of the length of the density interval where the match occurs and the length of the longer of the two density intervals.

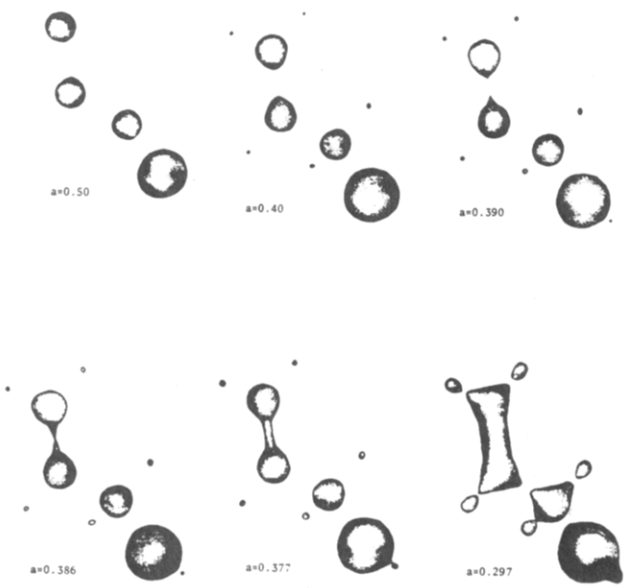

Figure 2. Selected families of density domains of the allyl alcohol molecule for high densities. Electron density values along isodensity contours are given in atomic units.

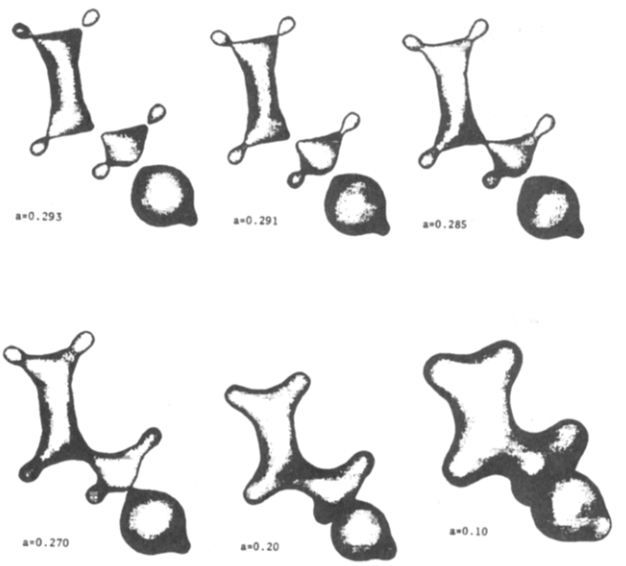

Figure 3. Selected families of density domains of the allyl alcohol molecule for low densities. Electron density values along isodensity contours are given in atomic units.

If a shape group analysis is performed on each DD, then the Betti numbers of these shape groups provide a numerical shape code based on the density domains. The usual shape codes are the vectors or the matrices of the one dimensional Betti numbers for the finite number of parameter intervals (a_i, a_{i+1}) and (b_j, b_{j+1}) for each DD, as described in the Introduction. These shape codes are ordered into vectors and matrices the same way as for ordinary shape groups, and any one of the usual methods of vector and matrix comparisons is applicable.

Our second example is that of the most stable conformation of allyl alcohol, CH_2=CH—CH_2—OH, as calculated with the GAUSSIAN 88 and GSHAPE 90 programs, using a 6-31G* basis set. This molecule, with many subtle shape features, represents a useful test case for topological shape analysis methods.[25] In contrast to water, this molecule is fully suitable to show a complexity of shape features expected in actual molecules of pharmacological significance.

A sequence of topologically different families of density domains of selected isodensity surfaces of CH_2=CH—CH_2—OH are shown in Figures 2 and 3. These DDs show many chemically interesting features. When considering a gradual decrease of the isodensity contour value a, the protonic DD that appears last (at $a = 0.390$) is that of the OH proton. Interestingly, this is the very protonic DD that loses its separate identity the earliest (at $a = 0.377$), and other protonic DDs join the DDs of neighboring carbon atoms at much lower electron densities (the first one at $a = 0.297$). Clearly, the electronegative oxygen lowers the electron density about the OH proton, hence this protonic DD appears last, but the oxygen also increases the electron density in the space between the two nuclei, hence this is the first bond formed that involves a proton. As expected, the double bonded pair of carbon DDs join first (at $a = 0.386$).

Especially interesting are the occurrence and persistence of certain DDs of the more common functional groups of organic chemistry. By far the most persistent DD is that of the OH group that preserves its separate identity within a very wide density range, from $a = 0.377$ to a value slightly beyond $a = 0.285$. The CH_2=CH vinyl group has its separate DD in the approximate density range between $a = 0.290$ and $a = 0.286$, whereas the CH_2 group has its own DD in the range between $a = 0.292$ and $a = 0.286$. The DD of the CH_2=CH—CH_2-

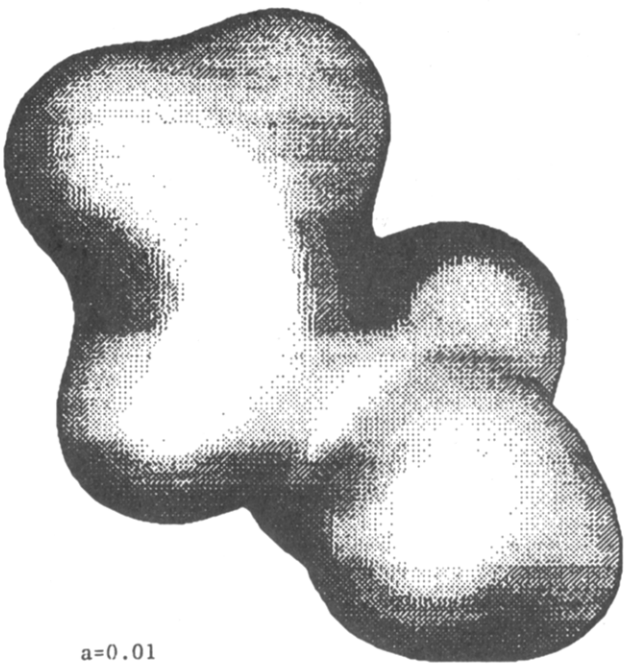

Figure 4. Single density domain isodensity contour of the allyl alcohol molecule for $a = 0.01$ atom unit, suitable for approximate representation of molecular size.

allyl group is rather prominent, it exists as a separate entity within the range of $a = 0.285$ and $a = 0.271$. At $a = 0.270$ the allyl alcohol molecule is bound together, by joining the DDs of the allyl and hydroxyl groups. At and below the density value $a = 0.270$ the molecule has only a single DD that is a topological sphere.

In Figure 4 the $a = 0.01$ isodensity contour surface of the allyl alcohol molecule is shown. This contour represents a formal molecular body with a volume that is suitable for representing the space requirement of the molecule in "average" intermolecular interactions. Three shape domain decompositions of this isodensity surface are shown in Figure 5, with respect to three different reference curvatures $b = 0.005$, $b = 0.000$, and $b = -0.008$ (that is, with respect to three different reference spheres, as described in the Introduction). The heavily shaded domains contain points where at most one local canonical curvature of the surface is less than the reference value b. The lightly shaded areas are the D_2 domains

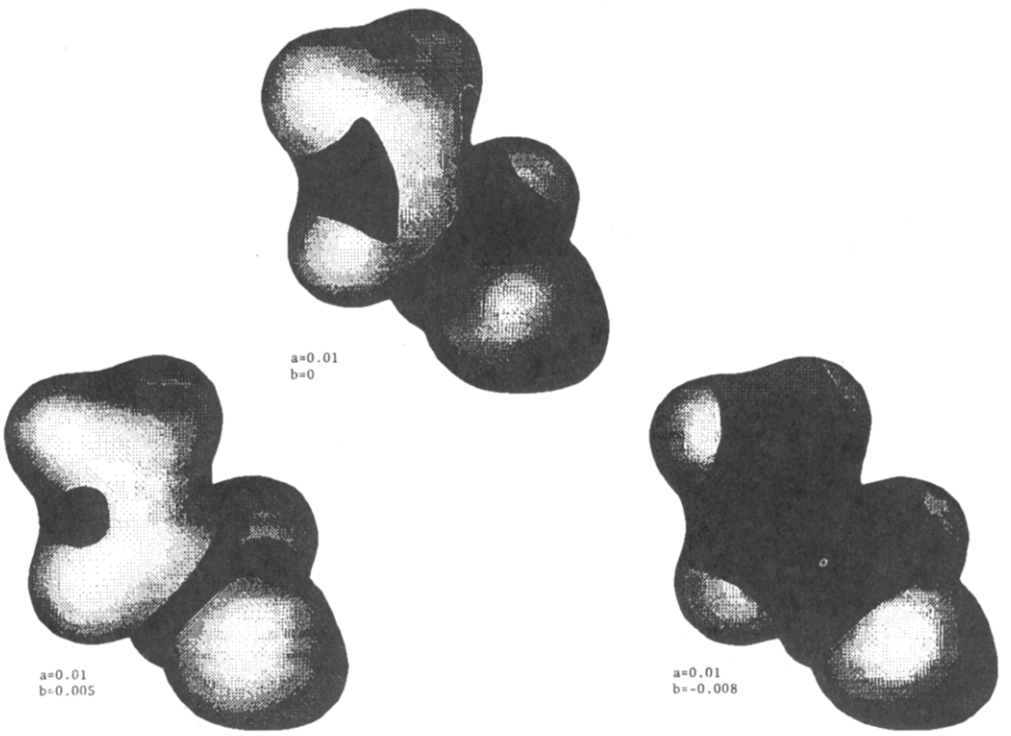

Figure 5. Shape domains of a single density domain isodensity contour of the allyl alcohol molecule for $a = 0.01$ atomic unit and for three reference curvature values, $b = 0.005$, 0.000, and –0.008.

relative to reference curvature b. Even with a small change in parameter b, the pattern of local shape domains changes considerably, leading to important topological differences. In all three cases a truncation of D_2 domains leave topologically different objects, hence these curvature domain patterns on the same isodensity surface lead to different shape groups for each of the three b values. It is important to realize, however, that for the whole range of possible b values there are only finitely many topologically different shape groups.

The DD approach also provides a topological tool for finding the range of molecular deformations which preserve the chemical identity of functional groups. It is an important question to decide whether a given deformed arrangement of a molecular fragment can or cannot be regarded as a functional group. In an earlier study[26] a geometrical criterion has been used to identify families of nuclear arrangements (domains in nuclear configuration space) where a given functional group occurs.

The DD approach provides a natural alternative criterion, formulated as follows. A chemical species of a given (possibly highly distorted) nuclear configuration K is regarded to contain a specified functional group f if there exist some electron density value a at which there is a DD that contains all the nuclei of f and if a gradual increase of the electron density value a leads to the same sequence of DD families as the sequence present for the equilibrium nuclear configuration of f in a typical molecule containing f. There is some ambiguity in the choice of a typical molecule that contains f. However, the sequence of families of density domains occurring along a density change is usually unaffected by the choice of molecules containing f, as long as the equilibrium configurations are considered for all.

T-HULL SHAPE ANALYSIS METHOD FOR FUSED SPHERE VAN DER WAALS SURFACES

An alternative shape analysis method and similarity measure is based on the T-hull method.[27] By contrast to the techniques described above, this method is applicable not only to smooth molecular surfaces, such as isodensity contours, but also to approximate molecular model surfaces with seams or sharp edges, fused-sphere van der Waals surfaces, as well as to discontinuous dot representations.

The concept of T-hull of objects[27] has been introduced as a generalization of the concept of convexity. Here we shall give only a simple, intuitive description of this concept as follows. Consider two 3D objects, A and T, and assume that T is large enough so A fits within the interior of T. Object T can be moved relative to A, and in some arrangements only some parts or no parts of A will fall within T. Fix the position of A and consider all possible arrangements of T which contain A. The intersection of all these arrangements of T is the T-hull of A.[27]

By an appropriate choice of object T, the shape of the T-hull is usually simpler than the shape of the original object A. This observation can be exploited in devising similarity measures, satisfying the general conditions of (P,W)-similarity.[11] If two objects A_1 and A_2 are rather different, their T-hulls are likely to be more similar. A shape group analysis can be carried out on the T-hulls instead of the original objects, and similarities that are not well revealed by direct shape analysis on the original objects may be detected by a shape analysis on their T-hulls. This approach allows one to carry out actual shape analysis using a fixed sequence of reference objects, T_1, T_2, ... T_t. The level of complexity of reference object T that leads to equivalent T-hulls (or in a more general framework, maximizes the similarity of the T-hulls) can be used to characterize the similarity of the original objects A_1 and A_2.

The simplification of shape achieved by generating T-hulls can be exploited in simplifying the shapes of fused sphere van der Waals surfaces (VDWS). These surfaces have artificial features at the seams of fused spheres, where the surfaces are not differentiable, showing large differences between two one-sided derivatives at the points of the seam. If the reference object T itself has a smooth surface, then the T-hull of a VDWS still may have lines along which differentiability is not assured, but the differences between two one-sided

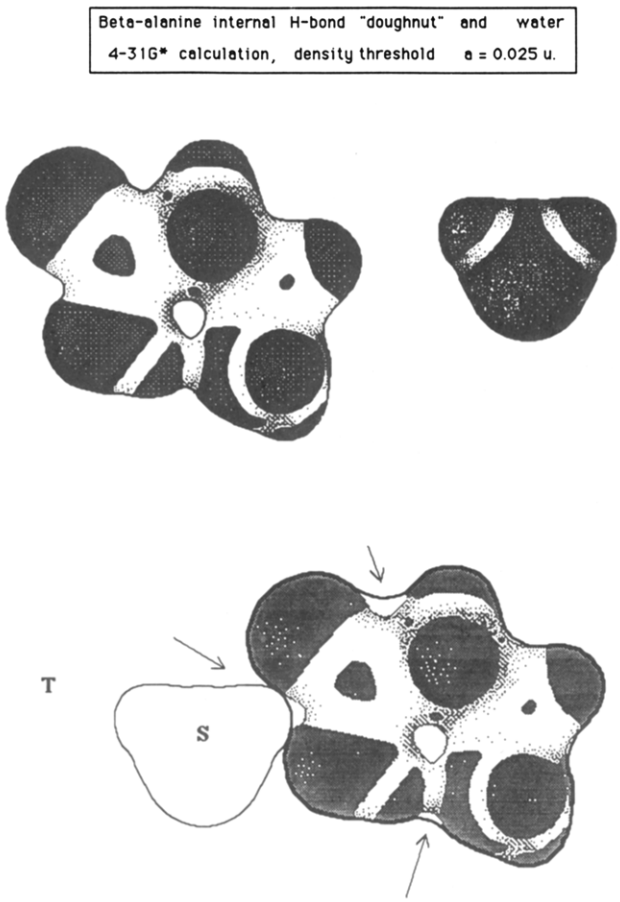

Figure 6. T-hull of an isodensity contour of a conformer of the β-alanine molecule, where T is the relative complement of a water isodensity contour body S, body at a density threshold value of a = 0.025, as calculated with a 4-31G* basis set. Dark areas represent D_2-type curvature domains.

derivatives at these points are usually reduced.

With a suitable choice of reference object T, the T-hull of a molecular contour surface G becomes the solvent contact surface (not to be confused with the solvent accessible surface) of the molecule. If S is a body enclosed by a contour surface of the solvent molecule, take T as the relative complement of S, that is, as the 3D space from where S is removed. The T-hull of G is the solvent contact surface of G. The construction of this surface is the simplest if S is spherical, since then no directional considerations apply.

An example of a T-hull of an isodensity contour of a conformer of the β-alanine molecule is shown in Figure 6. (A detailed shape analysis of the various stable conformers of this molecule will be presented elsewhere.[28]) The reference object T is chosen as the relative complement of a water isodensity contour surface and the body enclosed by it, denoted by S. Both contours are taken at a density threshold value of A = 0.025, as calculated with a 4-31G* basis set. The curvature domains on both molecular contour surfaces are also indicated, dark areas representing D_2-type domains. The solvent contact surface is shown by the heavy line in the lower part of the figure, with arrows indicating some of the locations where the solvent-generated surface deviates from the isodensity contour. The chosen density threshold value of 0.025 unit is only slightly lower than the unique $a_0(M_1,M_2)$ value for the given molecule pair M_1,M_2, where $a_0(M_1,M_2)$ is defined as the highest density threshold where the contact surface obtained with M_2 as the solvent and the isodensity contour of molecule M_1 are identical.

Note that the isodensity contour of the β-alanine molecule is a topological doughnut, due to the internal hydrogen bond between the OH proton and the N atom of the NH_2 group (lower part of the contour). By contrast, the solvent contact surface is a topological sphere. This example shows that major topological differences may exist between the isodensity and solvent contact surfaces, even if the actual density threshold a is in the vicinity of the $a_0(M_1,M_2)$ value.

CONCLUSIONS

A combination of the density domain approach and the shape group method leads to a shape characterization technique that provides a shape similarity measure and also a detailed description of chemical bonding. The T-hull method is proposed for a shape-similarity measure applicable to nonsmooth molecular surface representations, such as fused-sphere van der Waals surfaces, solvent accessible surfaces, and dot representations.

ACKNOWLEDGMENT

Research work leading to this report has been supported by both operating and strategic research grants from the Natural Sciences and Engineering Research Council of Canada.

REFERENCES AND NOTES

(1) Martin, Y. C. *Quantitative Drug Design: A Critical Introduction*; Dekker: New York, 1978.
(2) Richards, W. G. *Quantum Pharmacology*; Butterworths: London, 1983.
(3) Franke, R. *Theoretical Drug Design Methods*; Elsevier: Amsterdam, 1984.
(4) Dean, P. M. *Molecular Foundations of Drug–Receptor Interaction*; Cambridge University Press: New York, 1987.
(5) Mezey, P. G. Molecular Surfaces. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH Publishers: New York, 1990.
(6) Mezey, P. G. *Int. J. Quant. Chem. Quant. Biol. Symp.* **1986**, *12*, 113.
(7) Mezey, P. G. *J. Comput. Chem.* **1987**, *8*, 462.
(8) Mezey, P. G. *Int. J. Quant. Chem. Quant. Biol. Symp.* **1986**, *14*, 127.
(9) Mezey, P. G. *J. Math. Chem.* **1988**, *2*, 299.
(10) Mezey, P. G. *J. Math. Chem.* **1988**, *2*, 325.
(11) Mezey, P. G. Three-Dimensional Topological Aspects of Molecular Similarity. In *Concepts and Applications of Molecular Similarity*; Johnson, M. A., Maggiora, G. M., Eds.; Wiley: New York, 1990.
(12) Mezey, P. G. Topological Quantum Chemistry. In *Reports in Molecular Theory*; Náray-Szabó, G., Weinstein, H., Eds.; CRC Press: Boca Raton, FL, 1990.
(13) Mezey, P. G. Non-Visual Molecular Shape Analysis: Shape Changes in Electronic Excitations and Chemical Reactions. In *Computational Advances in Organic Chemistry (Molecular Structure and Reactivity)*; Ogretir, C., Csizmadia, I. G., Eds.; Kluwer Academic: Dordrecht, 1991.
(14) Mezey, P. G. Dynamic Shape Analysis of Biomolecules Using Topological Shape Codes. In *The Role of Computational Models and Theories in Biotechnology*; Bertran, J., Ed.; Kluwer Academic Publishers: Dordrecht, 1992.
(15) Mezey, P. G. *Shape in Chemistry: Introduction to Molecular Shape and Topology*; VCH Publishers: New York, 1992.
(16) Vick, J. *Homology Theory*; Academic Press: New York, 1973.
(17) Arteca, G. A.; Mezey, P. G. *Int. J. Quant. Chem. Quant. Biol. Symp.* **1987**, *14*, 133.
(18) Arteca, G. A.; Jammal, V. B.; Mezey, P. G.; Yadav, J. S.; Hermsmeier, M. A.; Gund, T. M. *J. Mol. Graph.* **1988**, *6*, 45.
(19) Arteca, G. A.; Jammal, V. B.; Mezey, P. G. *J. Comput. Chem.* **1988**, *9*, 608.
(20) Mezey, P. G. *J. Math. Chem.* **1991**, *7*, 39.
(21) Mezey, P. G. *IEEE Eng. Med. Bio. Soc.* **1989**, *11*, 1907.
(22) Berlin, Th. *J. Chem. Phys.* **1951**, *19*, 208.
(23) Frisch, M. J.; Head-Gordon, M.; Schlegel, H. B.; Raghavachari, K.; Binkley, J. S.; González, C.; DeFries, D. J.; Fox, D. J.; Whiteside, R. A.; Seeger, R.; Melius, C. F.; Baker, J.; Martin, R.; Kahn, L. R.; Stewart, J. J. P.; Fluder, E. M.; Topiol, S.; Pople, J. A. *GAUSSIAN 88*; Carnegie-Mellon Quantum Chemistry Publishing Unit: Pittsburgh, PA, 1988.
(24) Walker, P. D.; Arteca, G. A.; Mezey, P. G. *GSHAPE 90*; Mathematical Chemistry Research Unit, University of Saskatchewan: Saskatoon, Canada, 1990.
(25) Walker, P. D.; Mezey, P. G., unpublished data.
(26) Dubois, J.-E.; Mezey, P. G. *Int. J. Quant. Chem.* **1992**, *43*, 647.
(27) Mezey, P. G. *J. Math. Chem.* **1991**, *8*, 91.
(28) Heal, G.; Walker, P. D.; Ramek, M.; Mezey, P. G., unpublished data.