Shape Group Analysis of Molecular Similarity: Shape Similarity of Six-Membered Aromatic Ring Systems

P. Duane Walker,‡ Gerald M. Maggiora,† Mark A. Johnson,† James D. Petke,† and Paul G. Mezey*,‡

Mathematical Chemistry Research Unit, Department of Chemistry, University of Saskatchewan, Saskatoon, Saskatchewan, Canada, S7N 0W0, and Computer-Aided Drug Discovery, The Upjohn Company, Kalamazoo, Michigan 49007

Received January 12, 1995⊗

The nonvisual, algorithmic shape group method (SGM) is a technique for characterizing molecular shape that takes into account the entire electron density of molecules. The method provides a set of "shape codes" for each molecule, derived from a two-dimensional (a,b)-parameter map, where a represents the values of the electron density and b is related to the local curvature of molecular isodensity contour (MIDCO) surfaces. An (a,b)-parameter map provides a unique characterization of a molecule's shape and can be stored electronically in a "molecular shape database". Shape information can, thus, be retrieved rapidly from such a database and used in shape comparisons when needed. Actual shape comparisons are carried out by pairwise matching of the (a,b)-parameter maps, which yields a numerical measure of *molecular shape similarity*. There is no need to reconstruct molecular electron densities for each comparison once the (a,b)-parameter maps have been determined. Consequently, the method is simpler and less expensive than more conventional methods that compare the electron densities directly. Moreover, as the numerical shape-similarity measure quantifies similarity, it is expected to become a useful tool in computer-aided molecular engineering and drug design. In the current work, the SGM-based shape-similarity measure is used to study the similarities of 22 six-membered aromatic ring systems, whose electron densities are calculated at the AM1 level of semiempirical theory. The shape-similarities computed for these molecules provide a basis for interpreting a variety of chemically intuitive shape relationships as well as some interesting counter-intuitive relations.

INTRODUCTION

Similarities between molecular conformations[1−3] are usually interpreted in terms of the nuclear arrangements of the molecules.[4−7] In contrast, similarities between the shapes of molecular bodies[8] are better described by comparisons of the fuzzy electron densities surrounding the nuclei. Shape analysis of the electron density provides a faithful description of the local and global space requirements of molecules. The particular shape characteristics of molecules provide important clues concerning their interactions and reactivity; shape similarity and shape complementarity can be used for interpreting many chemical and biochemical processes. In modern drug design and in the emerging field of molecular engineering, numerical measures of shape similarity and shape complementarity will play an increasingly important role.[9−11] In the context of molecular wave functions, quantum similarity measures can be computed for pairs of molecules;[12−14] analogous measures can also be generated for cumulative and discrete similarity analysis of electrostatic potentials and fields.[15]

All of these methods rely on the principle of relative shape characterization: when analyzing shape, two or more molecules are compared. By contrast, molecular shape can also be characterized in absolute terms, without reference to another molecule, leading to an absolute shape description. An early example of a technique for absolute shape description and shape comparison of specific molecular surfaces is based on the methods of Fourier analysis.[16]

An alternative technique, applicable for the entire, three-dimensional distribution of the electron charge density $\varrho(\mathbf{r})$, is based on the shape group method (SGM) and the associated (a,b)-parameter maps,[17−20] which serve as *numerical shape codes*. The electronic charge densities are computed from *ab initio* or semiempirical wave functions, such as the AM1 technique of GAUSSIAN 90.[21] Since our long term goal is the shape analysis of large, biologically important molecules, in several conformations, in this study the inexpensive AM1 technique has been used for the computation of electronic densities. The shape codes are then evaluated with the GSHAPE 90 program.[22] These shape codes provide a means for evaluating the similarity between pairs of molecular shapes.

Numerical descriptors for absolute shape characterization and efficient shape similarity measures, applicable to large families of related molecules, are important in many fields of chemistry and biochemistry, including quantum pharmacology and computer-aided drug design. When thousands of molecules need to be compared, visual shape inspection is inadequate, and reproducible, numerical computer methodology is required. The SGM, with its corresponding (a,b)-parameter maps, fulfills this role. Below, is a brief introduction to the methodology; more details and background information can be found in ref 8.

* To whom all correspondence should be addressed. Also at Department of Mathematics and Statistics, University of Saskatchewan, Saskatoon, Saskatchewan, Canada, S7N 0W0.
† The Upjohn Company.
‡ University of Saskatchewan.
⊗ Abstract published in *Advance ACS Abstracts*, April 1, 1995.

0095-2338/95/1635-0568$09.00/0

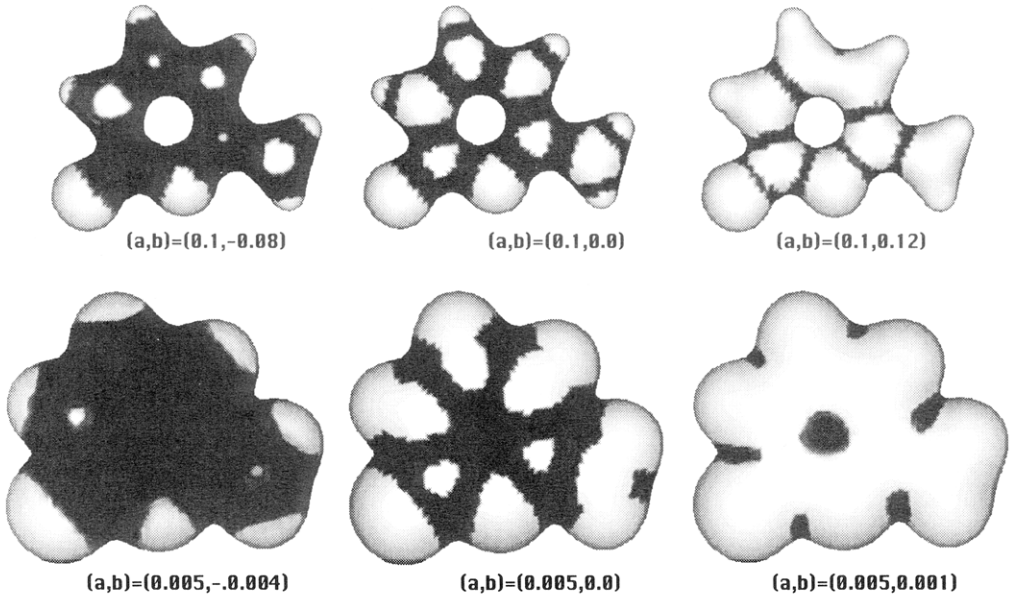

Figure 1. Three relative convexity domain partitionings of the $G(0.1)$ and $G(0.005)$ molecular isodensity contour (MIDCO) surfaces of the cytosine molecule, with respect to reference curvatures b_1, b_2, and b_3, respectively.

A BRIEF REVIEW OF THE SHAPE GROUP METHODS (SGM)

Shape group methods[17-19] are based on nonvisual, algebraic-topological, algorithmic shape characterization of fuzzy clouds of the electronic-charge density, or other properties, such as molecular electrostatic potential, that can be represented by three-dimensional functions. Molecular isodensity contours, or MIDCOs, are commonly used to model molecular surfaces.

For a fixed nuclear configuration K, a MIDCO surface $G(K,a)$, is the collection of all points of the 3D space where the electronic density $\varrho(\mathrm{K},\mathbf{r})$ is equal to the threshold value a

$$G(K,a) = \{\mathbf{r}:\ \varrho(K,\mathbf{r}) = a\} \qquad (1)$$

If only a single nuclear configuration K is considered and if this is obvious from the context, then the symbol K is suppressed in the notation.

Such MIDCOs can be used to describe formal molecular surfaces and approximate molecular size. Here, however, the focus is on the entire electronic-charge density, not only the peripheral regions of molecules. MIDCOs are still useful in this case as the entire electronic charge distribution of a molecule can be represented by a continuum of MIDCO surfaces. The SGM involves partitioning of each MIDCO into domains characterized, for example, by the local curvature properties of the surfaces. This step is followed by the formal removal of certain domains that satisfy a chosen criterion, for example, all those that are "more curved" than some reference curvature b. A *shape group* is the homology group (an algebraic-topological construct) defined on the remaining truncated surface.[17-19] Homology groups and shape groups can be introduced in a simple manner,[8] more details of homology theory can be found in refs 23 and 24.

Each surface $G(a)$ is subdivided into curvature domains $D_\mu(a,b)$: each point of the surface is assigned to one such domain, depending on how curved the surface is at the point when compared to the reference curvature b. In more precise terms, the local canonical curvatures (the eigenvalues of the local Hessian matrix defined for each point of G(a) over a local tangent plane) of each point of the surface are compared to the curvature of a reference tangent sphere of curvature b. The subscript μ is the number of eigenvalues which are less than the reference curvature b. As there are two eigen values for the local Hessian, the possible values for μ are 0, 1, and 2. If $b = 0$ (*i.e.*, if curvatures are compared to a "sphere" of zero curvature, that is, to a reference plane), then the values of 0, 1, and 2 refer to locally concave, saddle, and convex domains, respectively. By a generalization of the above concepts, a domain is said to be locally convex *relative to the reference curvature b* if both eigenvalues of the local Hessian matrix are less than b for all points in the domain.[17-19] One may visualize a *positive* reference curvature b as the curvature of a tangent sphere of radius $1/b$, brought into contact with the MIDCO from the *outside* of $G(a)$ (that is, densities $\varrho > a$ and the tangent sphere are on different sides of the local tangent plane). A *negative* reference curvature b indicates a formal tangent sphere of radius $1/|b|$, brought into contact with the MIDCO from the *inside* of $G(a)$ (that is, densities $\varrho > a$ and the tangent sphere are on the same side of the local tangent plane).

In Figure 1 three curvature domain partitionings of each of two MIDCOs, $G(0.1$ au) and $G(0.005$ au) of the cytosine molecule are shown (au = atomic unit), where the light areas correspond to the locally convex domains $D_2(a,b_1)$, $D_2(a,b_2)$, and $D_2(a,b_3)$, relative to several different reference curvatures, $b_1 = -0.08$, $b_2 = 0.0$, and $b_3 = 0.12$, for the $G(0.1)$ MIDCO, and $b_1 = -0.04$, $b_2 = 0.0$, and $b_3 = 0.001$, for the $G(0.005)$ MIDCO, respectively. Evidently, the patterns of shape domains change considerably by changes of the density threshold a and the reference curvature b. Nevertheless, the topology of these patterns remain invariant within small intervals of a and b, and there are *only a finite number of topologically different patterns* within the entire range of chemically relevant electron densities a and possible local curvatures b for MIDCOs. This observation is the basis for a topological characterization of molecular shapes,[17-19] that involves all possible MIDCOs and all possible reference curvatures, providing, indeed, a detailed shape characterization of the entire electronic density.[8] This technique follows the method developed for the shape characterization of

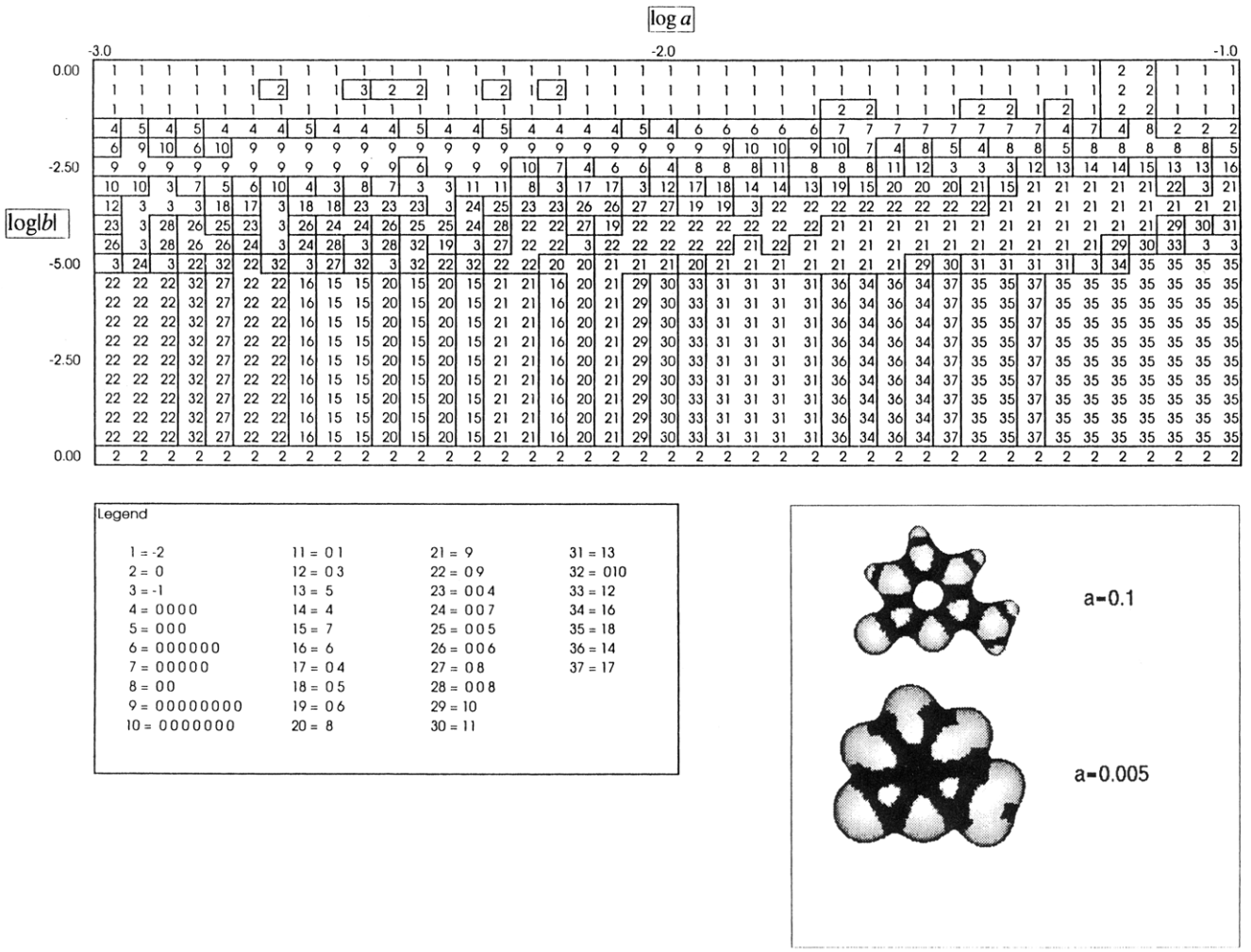

Figure 2. Two MIDCOs, $G(0.1)$ and $G(0.005)$, and the (a,b)-parameter map of cytosine.

potential energy hypersurfaces and patterns of reaction mechanisms.[7]

A *truncated surface* is obtained by removing the union of all domains of the same class μ from the isodensity surface

$$G_{\mu}(a,b) = G(a)\backslash \cup D_{\mu}(a,b) \qquad (2)$$

where symbols $\cup$ and $\backslash$ stand for union and set theoretical subtraction, respectively. This truncation transforms the shape analysis of a continuum of $G(a)$'s into a simple, discrete group-theoretical problem of algebraic homology.[23,24] In the simplest case, truncation results in a single surface with some holes. In other cases, truncation leads to fragmentation of the surface into a number of pieces. The truncated surfaces can consist of one or more disconnected surfaces with none, one, or more holes. In this work the truncations will always be applied to the D_2 domains (μ = 2, domains "locally convex relative to curvature b"), and the points in the D_1 and D_0 domains will be collected into single formal curvature domains which will be referred to as D^*_0 domains.

The algebraic homology groups of the truncated surface are the shape groups of the original MIDCO surface.[17-19] The Betti numbers of dimension zero, one, and two, are important shape group invariants. A one-dimensional Betti number (informally called a first Betti number, related to the number of holes in the surface) is the rank of the one-dimensional shape group for each piece of a truncated surface. The Betti numbers are used as shape descriptors and depend on the particular choice of electron density and curvature parameters, a and b, respectively. Another natural alternative is to take the entire family of pieces as a single object and the set of corresponding Betti numbers as shape descriptors,[8] although this approach provides somewhat less information.

A continuum of MIDCO surfaces $G(a)$ is characterized, for *all chemically important density thresholds a* and for all curvature domain partitionings according to *all relevant reference curvature values b*.

SHAPE CODES BASED UPON (a,b)-PARAMETER MAPS

The result of a shape group analysis is a family of Betti numbers, generated for all choices of the parameters a and b. In the application described in this study, we shall restrict the analysis to the one-dimensional Betti numbers. Note that the "pseudo Betti numbers" -2 and -1, have been introduced[20] to describe complete truncation (elimination, that is, a "no group" situation), and local ambiguity of the assignment (undecided at the given level of resolution), respectively (*vide infra*). The results for a given molecule can ultimately be represented as a two-dimensional (a,b)-parameter map, where each (a,b) pair corresponds to a specific location in the map, and the "value" assigned to that location is the corresponding set of Betti numbers. Regions having common sets of Betti numbers for all points within the

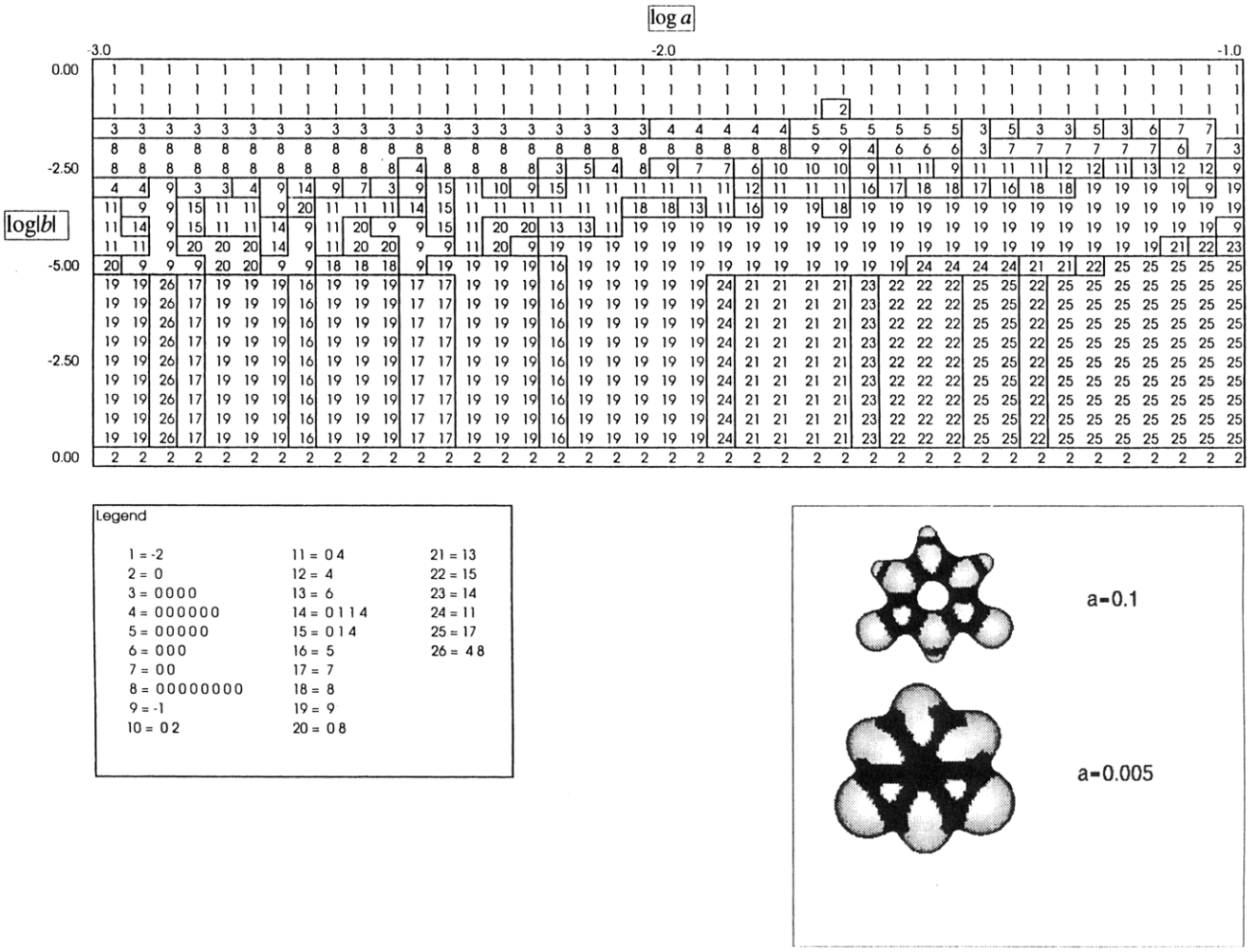

Figure 3. Two MIDCOs, $G(0.1)$ and $G(0.005)$, and the (a,b)-parameter map of uracil.

region can be viewed as "countries" of the map. This map contains detailed shape information for the range of density being considered. *Thus, the (a,b)-parameter map derived from an* SGM *analysis characterizes the shape of the entire range of* MIDCOs *for a given molecule.*

According to the approach followed in most of the shape group studies carried out to date, the set of Betti numbers, one number for each separate, truncated surface piece, is collected into a formal vector, which is taken as the shape descriptor. One technical problem with this approach is that there will be vectors of varying length for different values of a and b, depending on the number of surface pieces, making a uniform characterization cumbersome. Note that this problem does not arise with the alternative but less informative shape descriptors[8] mentioned above. To alleviate the problem with the current descriptors, the vectors of Betti numbers for a given molecule m can be encoded into a *unique shape integer* $L(m,a,b)$ using the following formula

$$L(m,a,b) = \prod_{i=0}^{M} (P_i)^{k(i)} \tag{3}$$

where $P_0 = 0$, and for $i \geq 1$, P_i is the ith prime number in the sequence 1, 2, 3, 5, 7, 11, ...; (where 1 is included as "honorary prime") each Betti number of value $(i - 2)$ is assigned to the prime number P_i; $k(i)$ is the number of disconnected surface pieces having their one-dimensional Betti number equal to $(i - 2)$; and where the value of M is two greater than the largest first Betti number for any of the disconnected pieces of the MIDCO $G(a)$. If the presence of one or more of the D_μ curvature domains is ambiguous at the given level of resolution, then a negative sign is assigned to $L(m,a,b)$.

The prime factorization theorem can then be used to decode $L(m,a,b)$, yielding the set of individual Betti numbers.

For example, if the truncation leaves five disconnected pieces, one having a Betti number of 3, another having a Betti number of 1, and the remainder having a Betti number of 0, then the shape integer $L(m,a,b)$ is 168.

$$L(m,a,b) = 2^3 * 3^1 * 5^0 * 7^1 = 168 \tag{4}$$

As a consequence of the prime factorization theorem, the Betti numbers 3, 1, and 0 can be recovered from $L(m,a,b)$.

The (a,b)-parameter map (*vide infra*) is a two-dimensional map containing the Betti number information for ranges of both the density parameter a and the curvature parameter b, in the form of the shape integers $L(m,a,b)$. Note that in the actual (a,b)-parameter maps of this study each $L(m,a,b)$ is labeled by a single small integer to simplify the representation of the maps. Also, when displaying these maps as matrices, it is convenient to replace the actual shape integers (which can become rather large) with the much smaller serial numbers of the shape integers in the order of their occurrence. The correspondence between the actual sets of Betti numbers and the small integer label of $L(m,a,b)$ is specified in a supplementary legend table.

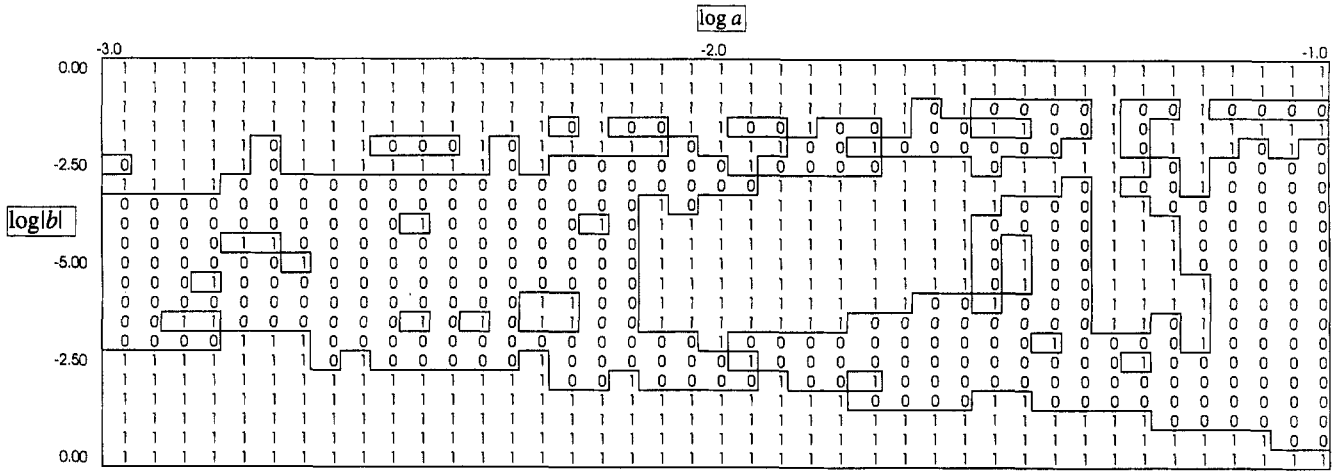

Figure 4. Binary similarity map ("match map") derived from the (a,b)-parameter maps of uracil **18** and cytosine **19**.

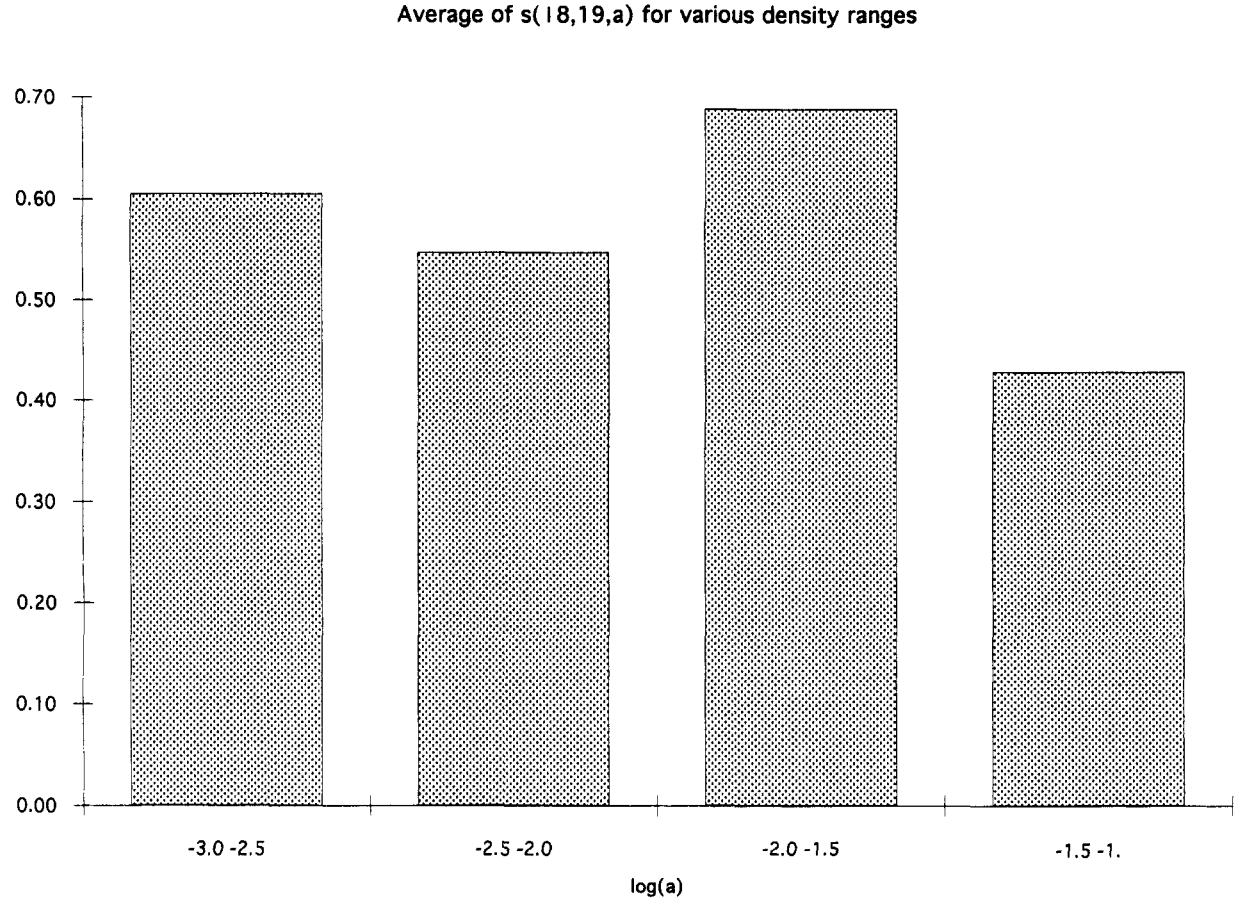

Figure 5. Plot of four (logarithmic) subranges, $[-3,-2.5]$, $[-2.5,-2.0]$, $[-2.0,-1.5]$, and $[-1.5,-1]$ of density dependent similarity measures for the cytosine−uracil pair, derived from the binary similarity map of Figure 4.

Due to the exponentially varying nature of the electronic density with the radial distance, and to the wide range of relevant curvatures, the scales for the (a,b)-parameter map are chosen to be logarithmic.

The resolution of these maps is standardized in this work to a grid with 41 grid points in density by 21 grid points in curvature. The range of density covered is from 0.001 to 0.1 au or in log units from −3.0 to −1.0. The map actually covers two ranges of the curvature parameter b, corresponding to both positive and negative reference curvatures, interpreted as curving "away" and curving "inward", respectively, with reference to the formal body enclosed by the MIDCO $G(a)$. Since b can take positive or negative values, the absolute value of b is needed for negative b values in order to use a logarithmic scale. The range of values for b are from 0 to −5.0 in log units for both the value of b (positive values) and the absolute value of b (negative values). For both ranges, the value of log $b = -5$ corresponds to $|b| = 10^{-5}$, approximated by $b = 0$, implying that all curvatures of $|b| \leq 10^{-5}$ are formally compressed to a line of the map. This approximation is sufficient for shape representation within the range of density being considered.

As implied by the actual grid, the shape integers are computed for 861 choices of a and b. These integers are collected into a single vector of 861 elements and defined as the shape code for the molecular charge distribution. The

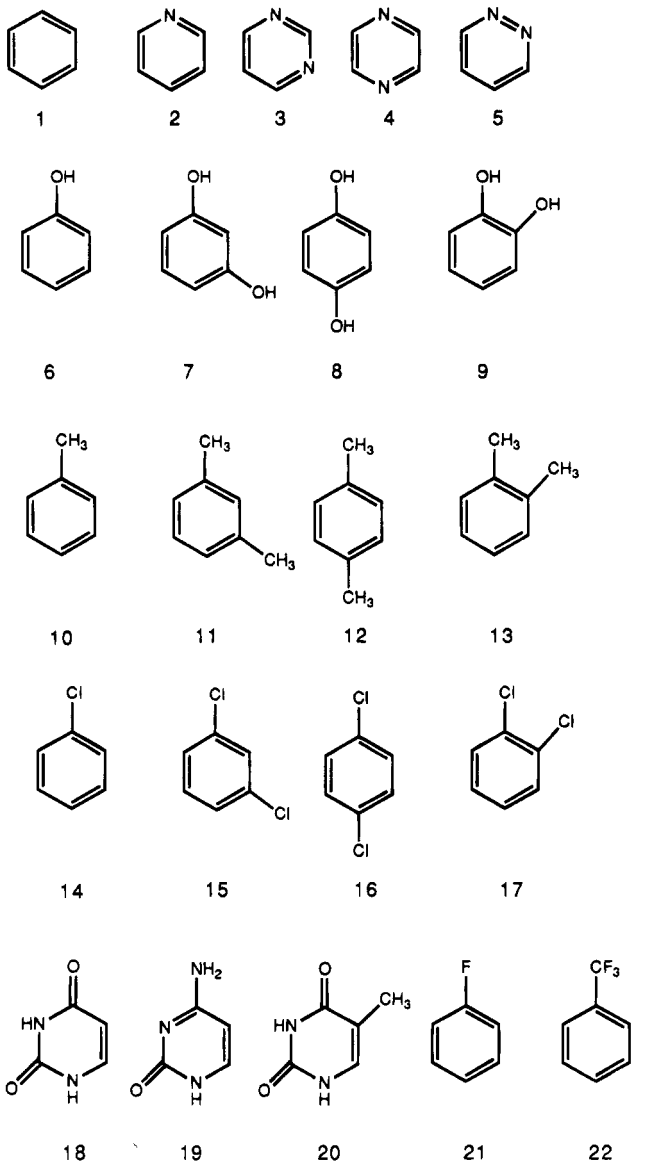

Figure 6. Structural formulas of the 22 aromatic molecules.

top row of the (a, b)-parameter map constitutes the first 41 elements of the shape code, and subsequent rows follow in descending order.

To illustrate these concepts, two examples of shape analysis, the resulting shape codes, and the (a,b)-parameter maps are shown in Figure 2 and 3, for cytosine and uracil, respectively. These molecules will be used to illustrate the concept of the binary similarity maps in the next section.

The shape codes can be stored in a database and retrieved as well as decoded for later comparisons. By using a standard grid size for all maps, a shape code of uniform length can be computed for all molecules.

MOLECULAR SIMILARITY MEASURES DERIVED FROM SHAPE CODES

The (a,b)-parameter maps can also be used for similarity analysis: the maps for two or more molecules can be compared numerically for each pair of (a,b)-parameter values, in order to assess and quantify molecular similarity.[8]

Following this approach, an alternative form of the shape code is introduced, where a list of the shape integers generates a vector of uniform dimension. This vector contains detailed shape information for the range of density being analyzed. Molecular similarity measures can be computed by comparing these shape codes for two or more molecules.

The calculation of a similarity measure for two molecular charge distributions is performed by comparing shape codes. By simply overlaying the two (a,b)-parameter maps, a new binary map is obtained: where the shape integers match, the numerical value of 1 is assigned, otherwise the value 0 is assigned to the corresponding location of the map. A simple shape-similarity measure is obtained by calculating the fraction of elements in the shape code which match for the chemical systems being compared.

For the ith molecule, the element of the (a,b)-parameter map at the a,b location is the shape integer $L(i,a,b)$. The binary difference map $\Delta(i,j,a,b)$ between molecules i and j is defined as

$$\begin{aligned}\Delta(i,j,a,b) &= 1 \quad \text{if} \quad L(i,a,b) = L(j,a,b) \\ &= 0 \quad \text{if} \quad L(i,a,b) \neq L(j,a,b)\end{aligned} \tag{5}$$

A (total) similarity measure is given as

$$S(i,j) = \sum_{k=1}^{N_a} \sum_{k'=1}^{N_b} \Delta(i,j,a_k,b_{k'})/(N_a \cdot N_b) \tag{6}$$

where N_a and N_b are the number of a and b values, respectively. One may wish to analyze similarity for specific electron density values, then the measure

$$s(i,j,a) = \sum_{k'=1}^{N_b} \Delta(i,j,a,b_{k'})/N_b \tag{7}$$

can be used. Clearly,

$$S(i,j) = \sum_{k=1}^{N_a} s(i,j,a_k)/N_a \tag{8}$$

A special measure applies if the reference curvature is restricted to the case of a tangent plane:

$$t(i,j) = \sum_{k=1}^{N_a} \Delta(i,j,a_k,0)/N_a \tag{9}$$

In one implementation of these measures, the points of the (a,b)-parameter maps which have ambiguous shape integer assignment at the given level of resolution of the grid, such as a mismatch that is one grid increment displaced from a match, are taken into account with a weight of 50%.

As an illustration, in Figure 4 the binary map of the uracil−cytosine pair is shown. There are numerous entries of 1 in this map, indicating a large number of matches between the Betti number families of the two molecules at various locations within the 41 × 21 grid.

In order to show the variation of similarity as a function of the electron density, the sum of eq 6 can be broken down to various subranges $[a_1,a_2]$ of density

$$S(i,j,[a_1,a_2]) = \sum_{[a_1,a_2]} \sum_{k'=1}^{N_b} \Delta(i,j,a_k,b_{k'})/(N_a \cdot N_b) \tag{10}$$

In Figure 5, a plot of four (logarithmic) subranges, [−3, −2.5], [−2.5,−2.0], [−2.0,−1.5], and [−1.5,−1] is shown for the cytosine−uracil pair. Clearly, the similarity is greater

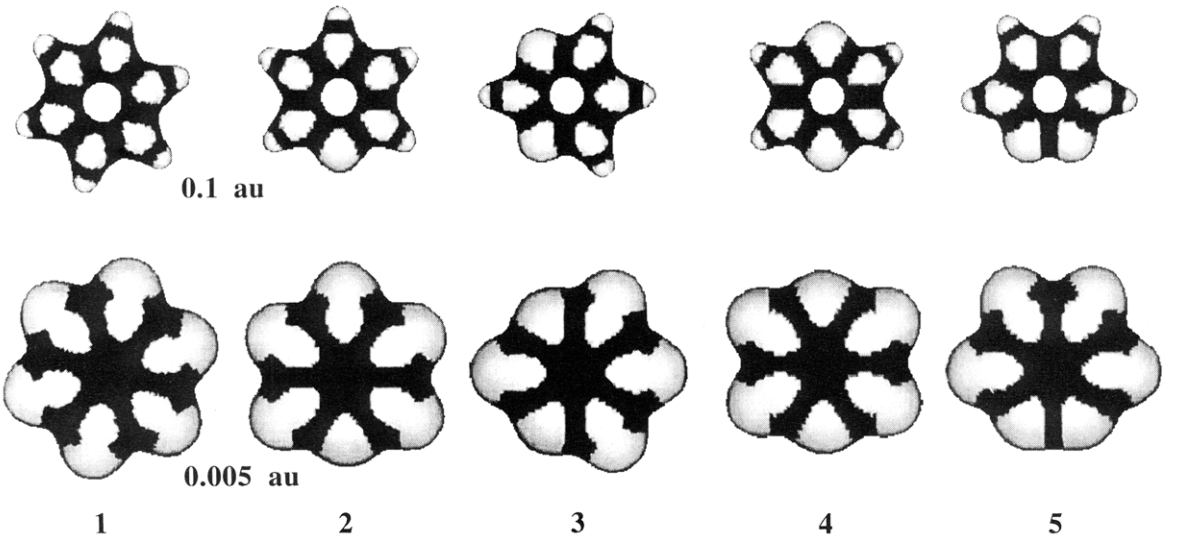

Figure 7. Two MIDCOs, $G(0.1)$ and $G(0.005)$, of molecules **1**−**5**, with shape domain patterns of reference curvature $b = 0$ indicated.

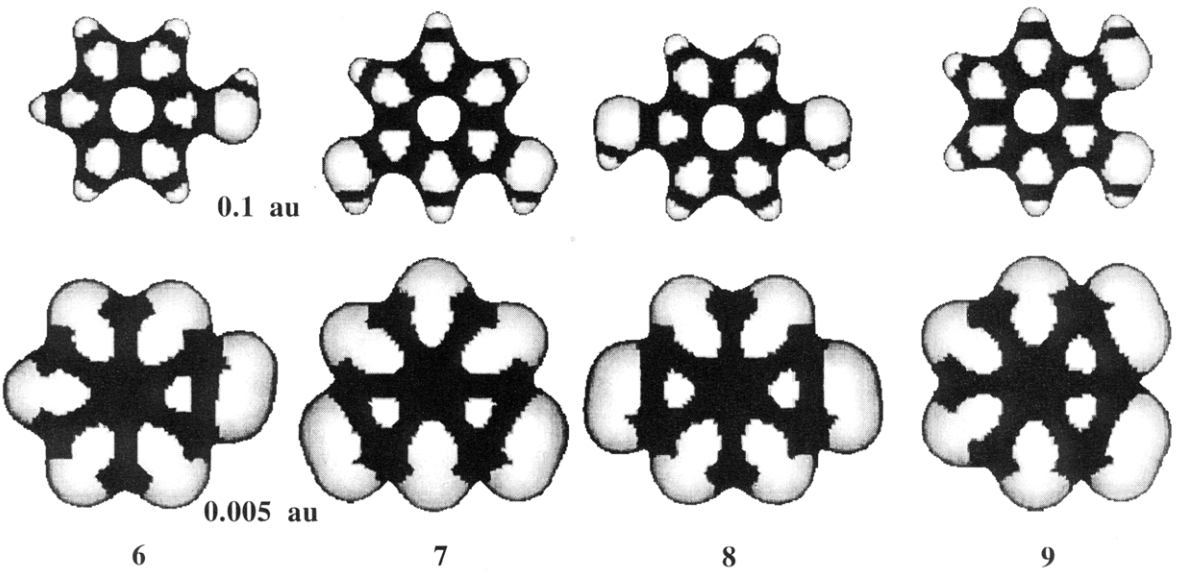

Figure 8. Two MIDCOs, $G(0.1)$ and $G(0.005)$, of molecules **6**−**9**, with shape domain patterns of reference curvature $b = 0$ indicated.

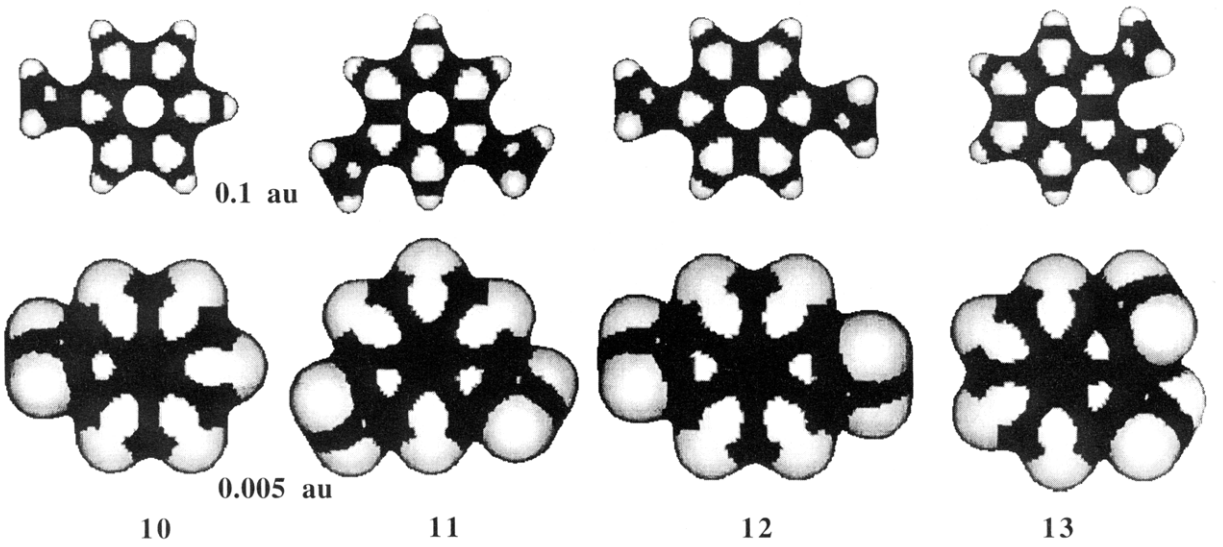

Figure 9. Two MIDCOs, $G(0.1)$ and $G(0.005)$, of molecules **1**−**13**, with shape domain patterns of reference curvature $b = 0$ indicated.

in the low density range ([−3,−2.5] on the logarithmic scale) than in the high density range ([−1.5,−1] on the logarithmic scale). At high density, the local nuclear neighborhoods are more different than the overall, bulky shapes at low density. Note, however, that we obtain a high level of similarity in the intermediate density range [−2.0,−1.5] (on the logarithmic scale), where the differences in the large scale features are already less important, and the differences in the local,

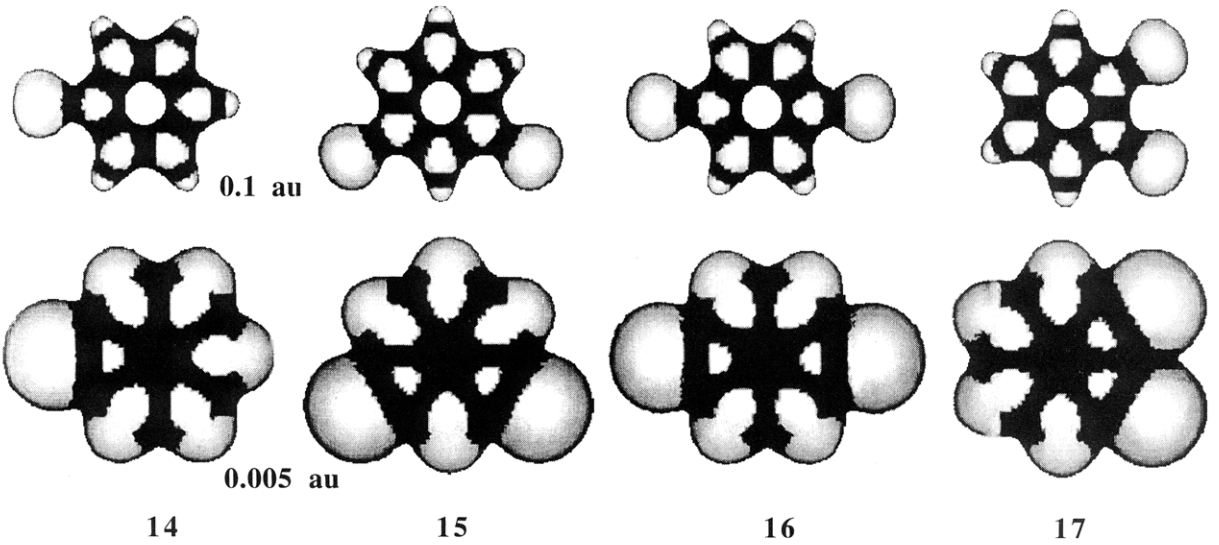

Figure 10. Two MIDCOs, $G(0.1)$ and $G(0.005)$, of molecules **14**−**17**, with shape domain patterns of reference curvature $b = 0$ indicated.

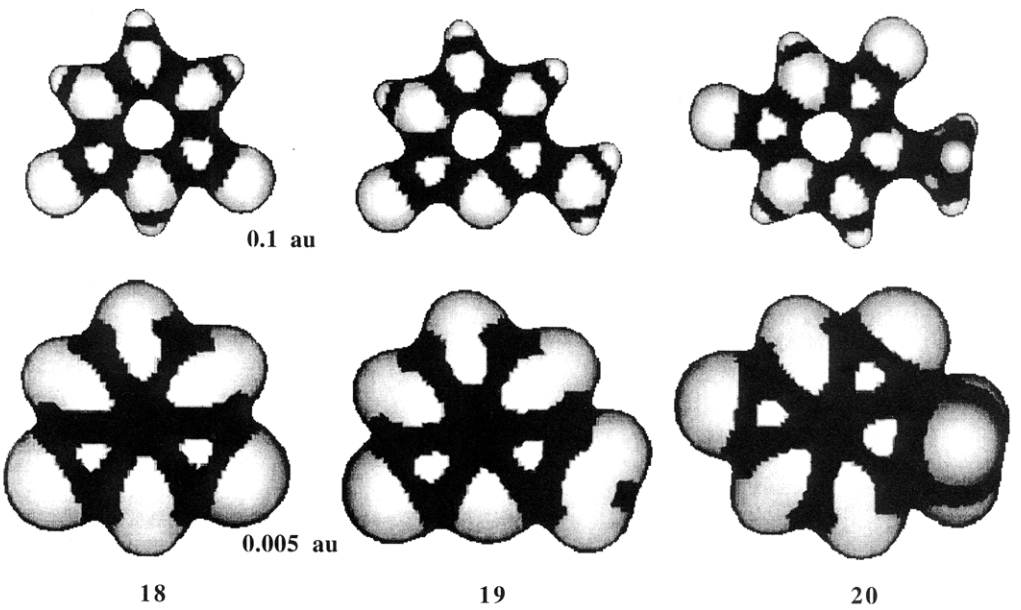

Figure 11. Two MIDCOs, $G(0.1)$ and $G(0.005)$, of molecules **18**−**20**, with shape domain patterns of reference curvature $b = 0$ indicated.

high density features are not yet dominant.

The general framework of the SGM is not restricted to electron densities. The SGM is applicable for the entire range of other molecular functions, such as electrostatic potential, spin density, or individual molecular orbitals, calculated by either semi-empirical or *ab initio* methods.[8,17−20]

SHAPE-BASED SIMILARITY CHARACTERIZATION OF A SERIES OF SIX-MEMBERED AROMATIC RING SYSTEMS

In this section, the application of (a,b)-parameter maps to the analysis of a series of aromatic molecules is described. Structural formulas of the 22 molecules studied are shown in Figure 6. The geometries of these molecules were optimized at semiempirical AM1 level using GAUSSIAN 90;[21] the shape codes and shape similarity measures were calculated using GSHAPE 90.[22]

Two, characteristic MIDCOs, $G(0.1)$ and $G(0.005)$, with shape domain patterns generated for reference curvature $b = 0$, are shown for these aromatic molecules in Figures 7−12.

Table 1 shows the computed similarities for all pairs of the 22 molecules displayed in Figure 6.

In general, the inferences we draw from these numbers relate to trends and are qualitative. Note that at the actual resolution of the (a,b)-parameter map grids, differences of 1% or less in the $S(i,j)$ similarity measures cannot be regarded as reliable. Any trend based on such small differences should be viewed with caution.

In Table 1 several trends are evident. As expected, the electronic density of benzene **1** shows the highest degree of shape similarity with the four unsubstituted nitrogen heterocycles, **2**, **3**, **4**, and **5**:

$$S(\mathbf{1},\mathbf{2}) = 0.71, \quad S(\mathbf{1},\mathbf{3}) = 0.70, \quad S(\mathbf{1},\mathbf{4}) = 0.73, \quad S(\mathbf{1},\mathbf{5}) = 0.69$$

Interestingly, the next group of molecules showing a high level of shape similarity with benzene are the chlorobenzenes **14**, **15**, **16**, and **17** as well as fluorobenzene **21**. The four phenols **6**, **7**, **8**, and **9** as well as the methyl-substituted benzenes **10**, **11**, **12**, and **13** show more complicated shape features at high densities due to the presence of the hydrogen

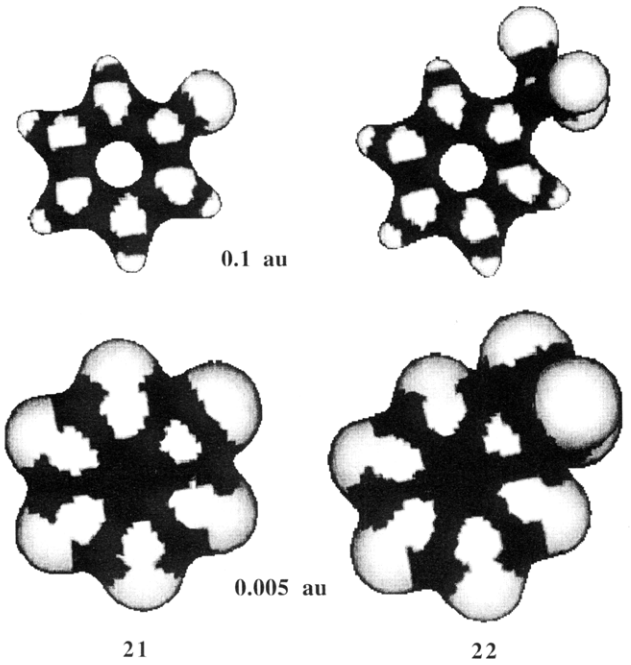

Figure 12. Two MIDCOs, $G(0.1)$ and $G(0.005)$, of molecules **21** and **22**, with shape domain patterns of reference curvature $b = 0$ indicated.

atoms in the substituents, and this lowers their similarity measures with benzene, when compared to the simpler halobenzenes. From the three pyrimidine bases **18**, **19**, and **20**, only **18** has shape-similarity greater than 0.5 with benzene; here, again, the presence of hydrogen atoms within the substituent groups (NH_2 and CH_3, in **19**, and **20**, respectively) lowers the level of shape similarity.

It is instructive to compare the groups of phenols, methyl substituted benzenes, and chlorobenzenes. For the monosubstituted compounds as well as for each of the *meta*, *para*, and *ortho* disubstituted benzenes, the highest shape similarity is found between the chlorobenzene and the corresponding phenol, whereas the methyl-substituted compounds show approximately the same level of dissimilarity to the corresponding phenol and chlorobenzene. The effect of the methyl protons renders the methyl-substituted compounds dissimilar to members of the other two families. This result is a consequence of shape features at high electron densities; if one focusses only on the low density, peripheral regions of the molecules, then a greater similarity is obtained between the methyl-substituted benzenes and the chlorobenzenes.

In Figure 13 the density dependence of similarity is shown for the toluene−chlorobenzene pair. Here, the only difference is between the Cl and Me groups, where the similarity is expected to decrease as the density threshold is increased. Indeed, the degree of similarity calculated for the four (logarithmic) ranges of electron density, [−3,−2.5], [−2.5,−2.0], [−2.0,−1.5], and [−1.5,−1], follows a monotonic trend, supporting the above expectation.

Whereas the high density features are the dominant cause of similarity orderings among the above molecular families, it is the shape of the low density ranges that dominates the dependence of similarities on the *placements* of substituents in the *meta*, *para*, and *ortho* benzenes. Within any group, the degree of pairwise similarity varies less than 4% among the *meta*, *para*, and *ortho* disubstituted benzenes. It is interesting to note that the *para* and *ortho* compounds do not show the highest similarity to one another; in fact, for the phenols and methylbenzenes the *para* and *ortho* pair shows the *lowest degree* of shape similarity. For the chlorobenzenes, the *similarity measures themselves are rather similar*; the actual value for the *meta*−*para* pair is the highest (0.89), whereas the similarity measures of the *ortho*−*para* and *ortho*−*meta* pairs are equal (0.87):

$$S(\mathbf{15},\mathbf{16}) = 0.89, \quad S(\mathbf{16},\mathbf{17}) = 0.87, \quad S(\mathbf{15},\mathbf{17}) = 0.87$$

This finding is explained by considering the opportunities for two substituents to form a single, bulky lump at low electron densities. This is least possible for the *para* compounds, easily accomplished by the *ortho* compounds at intermediate densities, and possible only at lower densities for the *meta* compounds. For these molecules, the prevailing shape similarities are dominated by features at low densities, as it is apparent from the binary similarity maps.

Entirely different conclusions are obtained when comparing the unsubstituted nitrogen heterocycles **3**, **4**, and **5**. In these molecules the highest measure of shape similarity, 0.92, is computed for the *para*−*ortho* pair, **4** and **5**,

Table 1. Similarity Measures Computed for All Pairs of the Set of 22 Molecules

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
1	1.00	0.71	0.70	0.73	0.69	0.45	0.43	0.46	0.45	0.45	0.43	0.46	0.44	0.62	0.60	0.63	0.61	0.60	0.50	0.44	0.67	0.42
2	0.71	1.00	0.82	0.76	0.73	0.42	0.45	0.43	0.42	0.42	0.40	0.41	0.41	0.49	0.44	0.44	0.45	0.53	0.47	0.44	0.50	0.40
3	0.70	0.82	1.00	0.87	0.87	0.44	0.44	0.44	0.41	0.42	0.41	0.42	0.42	0.47	0.46	0.47	0.45	0.49	0.45	0.42	0.49	0.38
4	0.73	0.76	0.87	1.00	0.92	0.42	0.43	0.42	0.40	0.41	0.41	0.41	0.41	0.43	0.45	0.45	0.44	0.49	0.44	0.42	0.46	0.37
5	0.69	0.73	0.87	0.92	1.00	0.41	0.43	0.42	0.40	0.40	0.41	0.42	0.42	0.44	0.46	0.48	0.44	0.50	0.44	0.44	0.46	0.39
6	0.45	0.42	0.44	0.42	0.41	1.00	0.43	0.51	0.44	0.53	0.40	0.41	0.40	0.70	0.44	0.47	0.47	0.53	0.63	0.48	0.66	0.48
7	0.43	0.45	0.44	0.43	0.43	0.43	1.00	0.86	0.86	0.65	0.51	0.52	0.51	0.43	0.70	0.65	0.63	0.59	0.54	0.57	0.44	0.75
8	0.46	0.43	0.44	0.42	0.42	0.51	0.86	1.00	0.82	0.63	0.49	0.52	0.49	0.46	0.70	0.71	0.68	0.65	0.60	0.59	0.43	0.77
9	0.45	0.42	0.41	0.40	0.40	0.44	0.86	0.82	1.00	0.63	0.50	0.54	0.53	0.42	0.65	0.61	0.66	0.58	0.54	0.61	0.44	0.75
10	0.45	0.42	0.42	0.41	0.40	0.53	0.65	0.63	0.63	1.00	0.44	0.44	0.45	0.52	0.52	0.51	0.53	0.52	0.49	0.56	0.53	0.58
11	0.43	0.40	0.41	0.41	0.41	0.40	0.51	0.49	0.50	0.44	1.00	0.81	0.83	0.40	0.52	0.50	0.48	0.47	0.43	0.52	0.41	0.45
12	0.46	0.41	0.42	0.41	0.42	0.41	0.52	0.52	0.54	0.44	0.81	1.00	0.80	0.43	0.55	0.50	0.50	0.47	0.45	0.53	0.45	0.48
13	0.44	0.41	0.42	0.41	0.42	0.40	0.51	0.49	0.53	0.45	0.83	0.80	1.00	0.42	0.52	0.48	0.50	0.46	0.44	0.54	0.43	0.47
14	0.62	0.49	0.47	0.43	0.44	0.70	0.43	0.46	0.42	0.52	0.40	0.43	0.42	1.00	0.59	0.62	0.62	0.62	0.54	0.44	0.85	0.45
15	0.60	0.44	0.46	0.45	0.46	0.44	0.70	0.70	0.65	0.52	0.52	0.55	0.52	0.59	1.00	0.89	0.87	0.69	0.59	0.51	0.58	0.64
16	0.63	0.44	0.47	0.45	0.48	0.47	0.65	0.71	0.61	0.51	0.50	0.50	0.48	0.62	0.89	1.00	0.87	0.76	0.61	0.46	0.59	0.60
17	0.61	0.45	0.45	0.44	0.44	0.47	0.63	0.68	0.66	0.53	0.48	0.50	0.50	0.62	0.87	0.87	1.00	0.73	0.63	0.48	0.58	0.64
18	0.60	0.53	0.49	0.49	0.50	0.53	0.59	0.65	0.58	0.52	0.47	0.47	0.46	0.62	0.69	0.76	0.73	1.00	0.71	0.45	0.58	0.59
19	0.50	0.47	0.45	0.44	0.44	0.63	0.54	0.60	0.54	0.49	0.43	0.45	0.44	0.54	0.59	0.61	0.63	0.71	1.00	0.48	0.47	0.58
20	0.44	0.44	0.42	0.42	0.44	0.48	0.57	0.59	0.61	0.56	0.52	0.53	0.54	0.44	0.51	0.46	0.48	0.45	0.48	1.00	0.45	0.57
21	0.67	0.50	0.49	0.46	0.46	0.66	0.44	0.43	0.44	0.53	0.41	0.45	0.43	0.85	0.58	0.59	0.58	0.58	0.47	0.45	1.00	0.45
22	0.42	0.40	0.38	0.37	0.39	0.48	0.75	0.77	0.75	0.58	0.45	0.48	0.47	0.45	0.64	0.60	0.64	0.59	0.58	0.57	0.45	1.00

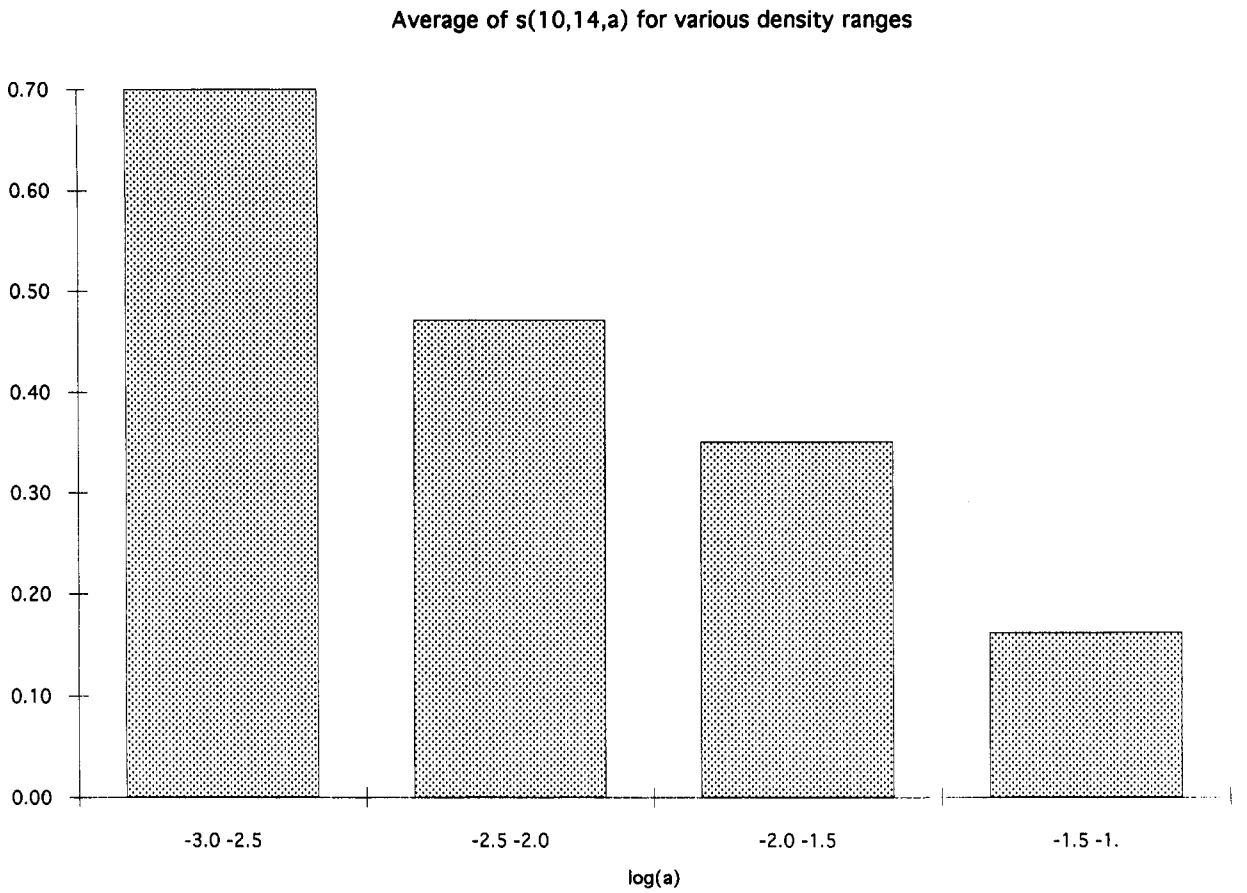

Figure 13. The density dependence of similarity is shown for the toluene−chlorobenzene pair. The *s(10,14,a)* quantity is plotted as a function of electron density *a* for four (logarithmic) subranges, [−3,−2.5], [−2.5,−2.0], [−2.0,−1.5], and [−1.5,−1].

$$S(\mathbf{4},\mathbf{5}) = 0.92$$

whereas for the other two combinations, **3**−**4**, and **3**−**5**, the measures are somewhat less, both 0.87,

$$S(\mathbf{3},\mathbf{4}) = 0.87, \quad S(\mathbf{3},\mathbf{5}) = 0.87$$

These similarity measures follow the chemical expectation based on reactivity properties of aromatic rings. Without substituents on the ring, the molecular shape is determined by the ring itself, influenced by the electronic pattern of alternation of heteroatoms within the ring.

In the group of the three pyrimidines bases **18**, **19**, and **20**, the presence of the methyl group renders **20** rather dissimilar from the other two molecules, whereas **18** and **19** show considerable similarity:

$$S(\mathbf{18},\mathbf{19}) = 0.71, \quad S(\mathbf{18},\mathbf{20}) = 0.45, \quad S(\mathbf{19},\mathbf{20}) = 0.48$$

The presence of a methyl group, with the three hydrogens providing some degree of additional shape complexity, significantly lowers the level of shape similarity.

If the hydrogens of a methyl group are replaced by fluorines, one may expect some similarities. The actual similarities, however, are less than expected. For the pair **10** and **22**, the similarity measure is only

$$S(\mathbf{10},\mathbf{22}) = 0.58$$

This low degree of shape similarity is caused by the prominence of the three bulky density lumps of the fluorines, that prevail at low densities where the much smaller hydrogen lumps are no longer distinguishable in the methyl group. The fluorines act as major substituents on their own, resulting in higher shape similarities with disubstituted benzenes **7**, **8**, and **9** as well as **15**, **16**, and **17**.

SUMMARY

Numerical shape similarity measures of molecular electron densities have been evaluated for a series of aromatic compounds, based on the SGM and the associated technique of (*a*,*b*)-parameter maps. The calculated shape similarity measures show some of the expected trends. For the unsubstituted nitrogen heterocycles the pair of the *ortho* and *para diaza* compounds exhibits the highest degree of shape similarity. However, for the substituted benzenes, where the shape is no longer determined exclusively by the aromatic ring, the *ortho*−*para* similarity is no longer the highest. In these compounds, the shape of the peripheral regions of the substituents, their internal details (presence of one or several hydrogens) as well as their merging into a single bulky lump (for *ortho* compounds, and to a lesser degree, for the *meta* compounds) are rather important. These factors result in higher similarities for pairs involving the *meta* compound.

We are currently investigating details of (*a*,*b*)-parameters maps as a means of clarifying the underlying electron-density features responsible for differences in the degrees of similarity. Numerical shape similarity analysis is expected to reveal further trends not immediately obvious from expectations based on experimental evidence, chemical intuition, or alternative theoretical considerations. We expect that shape similarity measures will become useful tools in computer-aided pharmaceutical drug design and molecular engineering.

ACKNOWLEDGMENT

This work was supported in part by an NSERC operating grant to P.G.M.

REFERENCES AND NOTES

(1) *Concepts and Applications of Molecular Similarity*; Johnson, M. A., Maggiora, G. M., Eds.; John Wiley & Sons, Inc.: New York, 1990.
(2) Johnson, M. A. *J. Math. Chem.* **1989**, *3*, 117.
(3) Nicholson, V. A.; Maggiora, G. M. *J. Math. Chem.* **1992**, *11*, 47.
(4) Eliel, E. L.; Allinger, N. L.; Angyal, S. J.; Morrison, G. A. *Conformation Analysis*; Wiley: New York, 1965.
(5) Mislow, K. *Introduction to Stereochemistry*; Benjamin: New York, 1966.
(6) Mezey, P. G. Topological Theory of Molecular Conformations In *Structure and Dynamics of Molecular Systems*; Daudel, R., Korb, J.-P., Lemaistre, J.-P., Maruani, J., Eds.; Reidel: Dordrecht, 1985.
(7) Mezey, P. G. *Potential Energy Hypersurfaces*; Elsevier: Amsterdam, 1987.
(8) Mezey, P. G. *Shape in Chemistry: An Introduction to Molecular Shape and Topology*; VCH Publishers: New York, 1993.
(9) Martin, Y. C. *Quantitative Drug Design: A Critical Introduction*; Dekker: New York, 1978.
(10) Richards, W. G. *Quantum Pharmacology*; Butterworths: London, 1983.
(11) Dean, P. M. *Molecular Foundations of Drug-Receptor Interaction*; Cambridge University Press: New York, 1987.
(12) Carbó, R.; Leyda, L.; Arnau, M. *Int. J. Quantum Chem.* **1980**, *17*, 1185.
(13) Hodgkin, E. E.; Richards, W. G. *Int. J. Quantum Chem., Quant. Biol. Symp.* **1987**, *14*, 105.
(14) Carbó, R., Calabuig, B. *Int. J. Quantum Chem.* **1992**, *42*, 1695.
(15) Petke, J. D. *J. Comput. Chem.* **1993**, *14*, 928.
(16) Leicester, S.; Bywater, R.; Finney, J. L. *J. Mol. Graph.* **1988**, *6*, 104.
(17) Mezey, P. G. *Int. J. Quantum Chem., Quant. Biol. Symp.* **1986**, *12*, 113.
(18) Mezey, P. G. *J. Comput. Chem.* **1987**, *8*, 462.
(19) Mezey, P. G. *J. Math. Chem.* **1988**, *2*, 325.
(20) Walker, P. D.; Arteca, G. A.; Mezey, P. G. *J. Comput. Chem.* **1993**, *14*, 1172.
(21) Frisch, M. J.; Head-Gordon, M.; Trucks, G. W.; Foresman, J. B.; Schlegel, H. B.; Raghavachari, K.; Robb, M. A.; Binkley, J. S.; Gonzalez, C.; Defrees, D. J.; Fox, D. J.; Whiteside, R. A.; Seeger, R.; Melius, C. F.; Baker, J.; Martin, R. L.; Kahn, L. R.; Stewart, J. J. P.; Topiol, S.; Pople, J. A. *GAUSSIAN 90*; Gaussian, Inc.: Pittsburgh, PA, 1990.
(22) Walker, P. D.; Arteca, G. A.; Mezey, P. G. *GSHAPE 90*; Mathematical Chemistry Research Unit: University of Saskatchewan, Saskatoon, SK, Canada, 1990.
(23) Spanier, E. H. *Algebraic Topology*; McGraw-Hill: New York, 1966.
(24) Greenberg, M. *Lectures on Algebraic Topology*; Benjamin: New York, 1967.

CI950006L