Iterated Similarity Sequences and Shape ID Numbers for Molecules

Paul G. Mezey

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

Received May 18, 1993*

Precise, computer-based description of shapes of complete molecular electron distributions is possible by the topological shape group method, SGM, developed earlier, using concise topological tools (shape groups) to describe the patterns of mutual arrangements of all possible curvature domains of all possible molecular isodensity contour surfaces (MIDCO's) of each conformer. The SGM gives a collective description of the shapes of all contours and patterns for all chemically relevant electron densities and curvatures. Conformational flexibility is accounted for by computing shape equivalence classes within conformational domains. The infinitely many individual arrangements are classified into a finite number of topologically different classes, and the classes are characterized by algebraic and numerical methods, using the principle of (P, W)-equivalence. This method reduces the complete molecular description to a much simpler model, while retaining all the essential shape information. The concept of (P,W)-equivalence is generalized for the treatment of *similarity sequences* and *iterated similarity*, important in comparisons in sequences of a large number of molecules where the criteria for similarity may change along the sequence. The full, 3D shapes of formal molecular bodies of electron distributions are described by precise, numerical shape codes calculated for series of MIDCO's obtained from quantum chemical calculations or from empirical, fused sphere models approximating actual MIDCO's. The shape codes are used for evaluating numerical shape similarity measures. In this contribution a family of *shape ID numbers*, leading to *shape ID vectors* is described, developed from earlier (a,b)-parameter maps of shape groups and shape matrices.

INTRODUCTION

Molecular similarity has been discussed in the context of the Hammond postulate (interrelating stable species and transition structures occurring along rection paths[1-9], with respect to various wave function similarity measures of Carbó and co-workers[10-14] and Richards and co-workers.[15,16] Systematic, comprehensive frameworks have been proposed for quantifying the degree of molecular similarity in chemistry [see, e.g. refs 17 and 18]. In some recent studies, the concepts of topology and fuzzy set theory have been applied for describing and quantifying molecular similarity.[19-29]

Whether two molecules are regarded as being similar or dissimilar is dependent on the context: similarity may refer to one or another particular type of molecular property or process. Topological methods are particularly suitable for the analysis of *shape similarity*.

Shape analysis methods fall into two classes, depending on whether the shape of an object is described with reference to its inherent shape properties or in comparison to the shape of another object. These two cases represent absolute and relative shape analysis, respectively.

In the strict sense, absolute shape analysis methods do not rely on similarity arguments. The (a,b)-parameter map of the shape groups of the electron density distribution of a molecule[19,23,30] and the various shape codes derived from it provide an *absolute shape characterization*. By considering these shape codes as vectors or matrices, they can be compared numerically. Once these shape codes are determined, molecular similarity can be evaluated by comparing shape codes and there is no need to re-evaluate these codes each time a molecule is compared to another. The *shape similarity measures of the first kind* are based on absolute shape descriptors.

By contrast, relative shape analysis and relative shape descriptors are based on comparisons. Such shape descriptors can change for each molecule, depending on the other molecule used for comparison. If n molecules are compared, then there are $n(n-1)/2$ molecule pairs, hence $n(n-1)/2$ families of relative shape descriptors of the given type. The *shape similarity measures of the second kind* are based on relative shape descriptors.

If a *shape representation P* of molecules is chosen, such as the electronic charge density, and if a topological *shape descriptor W* is selected for the characterization of P, for example, the shape groups or shape matrices of MIDCO's (molecular isodensity contours), then the concept of (P,W)-similarity[19] is interpreted by a topological equivalence, as follows:

The shapes of two molecules A and B are *(P,W)-similar*, [that is, *similar within the context* (P,W)], denoted by

$$\mathrm{A}\ (P,W)\ \mathrm{B} \tag{1}$$

if and only if for the actual (P_A,W_A) pair and a (P_B,W_B) pair

$$P_A = P_B \tag{2}$$

and

$$W_A\ \mathrm{h}\ W_B \tag{3}$$

hold, where h stands for the existence of a homeomorphic transformation between W_A and W_B.[19]

It is easily shown that the similarity relation A (P,W) B is an equivalence relation. The above framework of similarity analysis is the basis of the *GSTE Principle*: treating *geometrical similarity as topological equivalence*.[19]

ITERATED SIMILARITY

Similarity as used in everyday life does not always fulfill the transitivity requirement of equivalence relations. If the (P,W)-shape similarity relation is applied to a sequence of n objects, each object may appear similar to its immediate neighbors, but the two objects at the two ends of the sequence may appear as rather dissimilar.[30] If a whispered message is

* Abstract published in *Advance ACS Abstracts*, February 15, 1994.

0095-2338/94/1634-0244$04.50/0

sent through a line of people, the final message is often strikingly different from the initial one, even if the messages sent by two people who are neighbors in the line are likely to be similar. In this example, the actual criteria for similarity may gradually change along the line and the sequence may represent a whole range of similarities of a different nature.

The above problem can be treated within the general framework of (P,W)-shape similarity, if one allows the (P,W)-pair to change gradually along the sequence.[30] One may require only that there exists *some* (P,W)-pair that applies for each pair of messages that are neighbors in the sequence and that the (P,W)-pairs applied to pairs of messages not far from one another along the sequence are not too different.

In problems of the above type, we require that *similar* (P,W)-criteria, that is, *similar similarity criteria*, are to be used for pairs that are near one another along the sequence. Hence, our task is to assess the similarity of the (P,W)-criteria applied to the original objects (messages in the example). It is natural then to regard the similarity criteria themselves as a set of new objects to be compared and then to use the very same method for assessing their similarity. This scheme can be regarded as an *iterated similarity analysis*,[30] since some similarity criteria are applied to the very similarity criteria (P,W) used for comparing the n original objects, O_i, $i = 1,2, ..., n$.

The $n-1$, possibly different (P,W)-pairs applied to the n original objects O_i can be denoted by

$$(P^{(1)}{}_1,W^{(1)}{}_1), (P^{(1)}{}_2,W^{(1)}{}_2), ..., (P^{(1)}{}_{n-1},W^{(1)}{}_{n-1}) \quad (4)$$

The superscript (1) indicates that these are the similarity criteria of level one within the iterative scheme. The above sequence of $n-1$ criteria are taken *as a new set of objects*, and a new, higher level similarity criterion [denoted as $(P^{(2)},W^{(2)})$] is applicable to them. The $(P^{(2)},W^{(2)})$ similarity criterion does not have to remain constant when the similarity of $(P^{(1)}{}_1,W^{(1)}{}_1)$ to $(P^{(1)}{}_2,W^{(1)}{}_2)$ and the similarity of $(P^{(1)}{}_2,W^{(1)}{}_2)$ to $(P^{(1)}{}_3,W^{(1)}{}_3)$ are being judged, and, again, different criteria may be used. Further iterations are also possible, and the $n-2$ pairs of neighbors of the $(n-1)$-member sequence (4) can be judged by a sequence of $n-2$ similarity criteria

$$(P^{(2)}{}_1,W^{(2)}{}_1), (P^{(2)}{}_2,W^{(2)}{}_2), ..., (P^{(2)}{}_{n-2},W^{(2)}{}_{n-2}) \quad (5)$$

Here $(P^{(2)}{}_1,W^{(2)}{}_1)$ is used to compare $(P^{(1)}{}_1,W^{(1)}{}_1)$ to $(P^{(1)}{}_2,W^{(1)}{}_2)$, whereas $(P^{(2)}{}_2,W^{(2)}{}_2)$ is used to compare $(P^{(1)}{}_2,W^{(1)}{}_2)$ to $(P^{(1)}{}_3,W^{(1)}{}_3)$, and so on.

In general, one obtains the following iterative scheme:[30]

$$O_1, O_2, O_3, O_4, O_5, ..., O_{n-1}, O_n$$

$$(P^{(1)}{}_1,W^{(1)}{}_1),(P^{(1)}{}_2,W^{(1)}{}_2),(P^{(1)}{}_3,W^{(1)}{}_3),(P^{(1)}{}_4,W^{(1)}{}_4),...\ (P^{(1)}{}_{n-1},W^{(1)}{}_{n-1})$$

$$(P^{(2)}{}_1,W^{(2)}{}_1), (P^{(2)}{}_2,W^{(2)}{}_2), (P^{(2)}{}_3,W^{(2)}{}_3), ..., (P^{(2)}{}_{n-2},W^{(2)}{}_{n-2})$$

$$(P^{(3)}{}_1,W^{(3)}{}_1), (P^{(3)}{}_2,W^{(3)}{}_2), ..., (P^{(3)}{}_{n-3},W^{(3)}{}_{n-3})$$

$$\ll ...$$

$$(P^{(k-1)}{}_1,W^{(k-1)}{}_1), ..., (P^{(k-1)}{}_{n-k+1},W^{(k-1)}{}_{n-k+1})$$

$$(P^{(k)}{}_1,W^{(k)}{}_1), ..., (P^{(k)}{}_{n-k},W^{(k)}{}_{n-k})$$

$$\ll ... \quad (6)$$

The original objects O_i ($i = 1, 2, ..., n$) are regarded as the sequence at level zero, $j = 0$. On the jth level ($j > 0$), there are at most $n-j$ different similarity criteria $(P^{(j)}{}_i,W^{(j)}{}_i)$, $1 \le i \le n-j$. At this level, the criterion $(P^{(j)}{}_i,W^{(j)}{}_i)$ is used to compare the neighboring members $(P^{(j-1)}{}_i,W^{(j-1)}{}_i)$ and $(P^{(j-1)}{}_{i+1},W^{(j-1)}{}_{i+1})$ of the $(n-j+1)$-member sequence on the previous level $j-1$.

It is possible that on some level k a single criterion is sufficient. The first level k where a *common* criterion $(P^{(k)},W^{(k)}) = (P^{(k)}{}_1,W^{(k)}{}_1 = ... = (P^{(k)}{}_i,W^{(k)}{}_i) = ... = (P^{(k)}{}_{n-k},W^{(k)}{}_{n-k})$ is already applicable for the entire previous sequence $(P^{(k-1)}{}_1,W^{(k-1)}{}_1)$, $(P^{(k-1)}{}_2,W^{(k-1)}{}_2)$, ..., $(P^{(k-1)}{}_{n-k},W^{(k-1)}{}_{n-k})$ at level $k-1$ is called the factorial level, by virtue of an analogy with the *factorial level* of difference sequences of powers of integers.

The relevant result on difference sequences is of some interest in number theory. Take the sequence

$$d^{(0)}{}_0 = 0^k, d^{(0)}{}_1 = 1^k, d^{(0)}{}_2 = 2^k, d^{(0)}{}_3 = 3^k, d^{(0)}{}_4 = 4^k, d^{(0)}{}_5 = 5^k, ... \quad (7)$$

of the kth powers of integers. The kth difference sequence turns out to be the constant $k!$

Difference sequences are defined iteratively, for example, the ith element $d^{(j)}{}_i$ of the jth sequence is

$$d^{(j)}{}_i = d^{(j-1)}{}_{i+1} - d^{(j-1)}{}_i \quad (8)$$

We shall use the example of the second powers of integers. The sequence 0, 1, 4, 9, 16, 25, ... of squares of integers 0, 1, 2, 3, 4, 5, ... has the first difference sequence 1, 3, 5, 7, 9, ..., and the second difference sequence (the sequence of differences between subsequent elements of the first difference sequence) is 2, 2, 2, 2, ... Indeed, for the second powers the second difference sequence is a constant and is equal to $2! = 2$. In general, in the kth difference sequence one obtains a constant, the factorial $k!$ of the exponent k of the kth powers of integers. In Chart I, the scheme of the fourth powers is shown in detail. As expected, in the fourth difference sequence ($k = 4$) the constant difference $4! = 24$ is obtained.

Chart I

	$0^4, 1^4, 2^4, 3^4, 4^4, 5^4, 6^4, 7^4, 8^4, 9^4, 10^4, ...$
	0, 1, 16, 81, 256, 625, 1296, 2401, 4096, 6561, 10000, ...
($k = 1$)	1, 15, 65, 175, 369, 671, 1105, 1695, 2465, 3439, ...
($k = 2$)	14, 50, 110, 194, 302, 434, 590, 770, 974, 1202, ...
($k = 3$)	36, 60, 84, 108, 132, 156, 180, 204, 228, ...
($k = 4$)	24, 24, 24, 24, 24, 24, 24, 24, 24, ... (9)

Differences between numbers can be regarded as a similarity criterion. By application of this analogy to iterated similarity analysis, the serial number k of the first level where a common similarity criterion $(P^{(k)},W^{(k)})$ is already applicable is called the *power of the similarity sequence*. The final condition $(P^{(k)},W^{(k)})$ is called the *factorial similarity condition*.

The above system of similarity criteria provides a tool for topological similarity analysis for molecular sequences.

MOLECULAR SHAPE ID NUMBERS AND SHAPE ID VECTORS

The shape groups distributed along an (a,b)-parameter map of a molecule can be characterized by their Betti numbers. The entire map can be represented by a sequence of numbers ordered into a matrix or a vector.[25] In some applications, the process of surface truncation involved in the generation of the shape groups leads to several disjoint surface pieces. The shapes of these surface pieces can be characterized separately.

For each given (a,b)-pair, a size ordering of the surface pieces implies an ordering of their one-dimensional Betti numbers (informally called their "first" Betti numbers) into a sequence:

$$B(1), B(2), ..., B(k), ..., B(m) \quad (10)$$

It is common that the surface piece with the largest Betti number is the one with the largest surface area, implying that the sequence of Betti numbers in the above ordering is approximately the same as the *decreasing sequence* of the Betti numbers.

Within an (a,b)-parameter map, for different (a,b)-pairs both the number m of Betti numbers and the actual value of the Betti numbers may be different. As a result, direct coding methods relying on simple listings of the sequences (10) for each selected (a,b)-pair are not always uniform, and lists of greatly varying lengths are to be coded. In order to develop a shape ID number for each essentially different molecular shape, it is advantageous to use coding methods which are uniform for all (a,b)-pairs. One such method is described below.

A single number $c'(a,b)$ can be used to store the information on the entire sequence (10) of ordered Betti numbers for each parameter pair (a,b). This coding–decoding method relies on the prime factorization of integers.

Let p_i denote the ith prime number in the sequence

$$1, 2, 3, 5, 7, 11, 13, 17, ... \quad (11)$$

of primes. The code $c'(a,b)$ is defined as the following product:

$$c'(a,b) = 2^{B(1)+1} \times 3^{B(2)+1} \times ... \times (p_{k+1})^{B(k)+1} \times ... \times (p_{m+1})^{B(m)+1} \quad (12)$$

The number $c'(a,b)$ contains all information present in the values and the ordering of the original sequence of Betti numbers.[30] Using the prime factorization theorem, the code $c'(a,b)$ can be decoded easily: $c'(a,b)$ has a unique representation as a product of primes

$$c'(a,b) = 2^{r(2)} \times 3^{r(3)} \times ... \times (p_{k+1})^{r(k+1)} \times ... \times (p_{m+1})^{r(m+1)} \quad (13)$$

Here, the original Betti numbers can be computed from the exponents $r(k+1)$ by the following simple relation:

$$B(k) = r(k+1) - 1 \quad (14)$$

Below we shall consider an example. Let us assume that the truncated contour surface of a MIDCO falls into seven pieces at a given electron density threshold a and reference curvature b, and the shape group analysis gives the sequence

$$6, 3, 3, 0, 0, 0, 0 \quad (15)$$

of Betti numbers, ordered according to the decreasing size of the areas of the surface pieces. For this example, the code is

$$c'(a,b) = 2^{6+1} \times 3^{3+1} \times 5^{3+1} \times 7^{0+1} \times 11^{0+1} \times 13^{0+1} \times 17^{0+1} = 110\,270\,160\,000 \quad (16)$$

This single integer, representing shape information for the given (a,b)-parameter pair, is assigned to the associated (a,b)-point of the (a,b)-parameter map.

When the actual Betti numbers are needed, this integer $c'(a,b)$ can be decoded. The unique prime factorization of the number 110 270 160 000 gives

$$110\,270\,160\,000 = 2^7 \times 3^4 \times 5^4 \times 7 \times 11 \times 13 \times 17 \quad (17)$$

Using the relation $B(k) = r(k+1) - 1$ and the actual values of the exponents $r(k+1)$, the Betti numbers, as well as their ordering, are easily calculated. For example, the number 7 is the fifth prime number,

$$7 = p_5 \quad (18)$$

and the prime factor of 7 occurs on the first power in the number 110 270 160 000. Consequently,

$$r(5) = 1 \quad (19)$$

The above result implies that the surface piece of serial number $5 - 1 = 4$ according to decreasing size has a Betti number equal to $1 - 1 = 0$,

$$B(4) = r(5) - 1 = 1 - 1 = 0 \quad (20)$$

Whereas the above method encodes some size information in addition to shape, there is a price to pay: the method leads to large numbers for the code $c'(a,b)$, and for large numbers, the prime factorization can become very time consuming.

Some reduction in the magnitude of the code numbers is obtained in an alternative coding method where the size information is disregarded.[31] The Betti numbers are ordered into a sequence

$$B(1), B(2), ..., B(j), ..., B(m) \quad (21)$$

according to their *increasing* magnitude,[31] without consideration of the size of the surface pieces they represent. The code $c(a,b)$ is defined as the following product:

$$c(a,b) = p_{B(1)+2} p_{B(2)+2} \cdots p_{B(j)+2} \cdots p_{B(m)+2} \quad (22)$$

If a given value for Betti number $B(j)$ occurs t times in the sequence, then it is represented by a factor $(p_{B(j)+2})^t$, and the code $c(a,b)$ can be taken as the product of all these factors.

The decoding of $c(a,b)$ also relies on the prime factorization theorem; the original set of Betti numbers can be calculated from the prime factors of the number $c(a,b)$:

$$c(a,b) = (p_{i(1)})^{t(1)} (p_{i(2)})^{t(2)} ... (p_{i(s)})^{t(s)} ... (p_{i(w)})^{t(w)} \quad (23)$$

The Betti numbers which occur are

$$B(j) = i(s) - 2 \quad (24)$$

where

$$t(1) + t(2) + ... + t(s-1) < j \le t(1) + t(2) + ... + t(s-1) + t(s) \quad (25)$$

For the sequence

$$0, 0, 0, 0, 3, 3, 6 \quad (26)$$

considered above, the code $c(a,b)$ according to definition 22 gives

$$c(a,b) = 2^4 \times 7^2 \times 17 = 13\,328 \quad (27)$$

This number is much smaller than the number 110 270 160 000 of the $c'(a,b)$ code; consequently, the prime factorization in the decoding step takes much less time. Note, however, this code does not contain direct size information, although the approximate parallel trend in the magnitudes of the Betti numbers and surface areas provides some guideline. The sequence 0, 0, 0, 0, 3, 3, 6 of Betti numbers can be recovered easily from the unique prime factors of the number 13 328.

Both of the above two coding techniques compress several integers into a single integer number that can be decoded by prime factorization. In both cases, a uniform method is applicable to all choices of a and b values, allowing large variations in the length of the sequence of the Betti numbers.

The numerical value of either code, $c(a,b)$ or $c'(a,b)$, can be assigned to the (a,b)-location of the parameter map (a,b), and can be regarded as a shape ID number for the (a,b)-location of the parameter map. A list of these code values forms a vector, providing a numerical shape code for the entire (a,b)-map, for all relevant electron density values a and test curvature values b.

In practice,[31] a 41×21 grid is considered on the (a,b)-map, covering a range of [0.001–0.1 au (au = atomic unit)] of density threshold values a, and a curvature range of [(−1)–1.0] for the test spheres against which the local curvatures of the MIDCO are compared. The actual shape code of the entire 3D electron density is taken as the resulting 41×21 matrix C of integers, that can also be stored as an integer vector **C** of 861 components. This vector **C** can be taken as a "shape ID vector" of the molecule.

SUMMARY

A generalization of the concept of (P,W)-equivalence is presented, suitable for comparisons in sequences of large numbers of molecules where the criteria for similarity may change along the sequence. This generalization leads to the concept of iterated similarity. A family of shape ID numbers and shape ID vectors is developed from earlier (a,b)-parameter maps of shape groups.

REFERENCES AND NOTES

(1) Hammond, G. S. *J. Am. Chem. Soc.* **1955**, *77*, 334.
(2) Melander, L. *The Transition State*; Royal Society Special Publications: London, 1962; Vol. 16.
(3) Polanyi, J. C. *J. Chem. Phys.*, **1959**, *31*, 1338.
(4) Mok, M. H.; Polanyi, J. C. *J. Chem. Phys.*, **1969**, *51*, 1451.
(5) Miller, A. R. *J. Am. Chem. Soc.*, **1978**, *100*, 1984.
(6) Agmon, N. *J. Chem. Soc., Faraday Trans. 2* **1978**, *74*, 388.
(7) Murdoch, J. R. *J. Am. Chem. Soc.* **1983**, *105*, 2667.
(8) Arteca, G. A., Mezey, P. G. *J. Phys. Chem.* **1989**, *93*, 4746.
(9) Arteca, G. A.; Mezey, P. G. *J. Comput. Chem.* **1988**, *9*, 728.
(10) Carbó, R.; Leyda, L.; Arnau, M. *Int. J. Quantum Chem.* **1980**, *17*, 1185.
(11) Carbó, R.; Domingo, L. *Int. J. Quantum Chem.* **1987**, *32*, 517.
(12) Carbó, R.; Calabuig, B. *Comput. Phys. Commun.* **1989**, *55*, 117.
(13) Carbó, R.; Calabuig, B. *Int. J. Quantum Chem.* **1992**, *42*, 1681.
(14) Carbó, R.; Calabuig, B. *Int. J. Quantum Chem.* **1992**, *42*, 1695.
(15) Hodgkin, E. E.; Richards, W. G. *Int. J. Quantum Chem.* **1987**, *14*, 105.
(16) Burt, C.; Richards, W. G.; Huxley, P. *J. Comput. Chem.* **1990**, *11*, 1139.
(17) Johnson, M. A. *J. Math. Chem.* **1989**, *3*, 117.
(18) Johnson, M. A., Maggiora, G. M., Eds. *Concepts and Applications of Molecular Similarity*; Wiley: New York, 1990.
(19) 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.
(20) Mezey, P. G. *Int. J. Quantum Chem. Quant. Biol. Symp.* **1986**, *12*, 113.
(21) Mezey, P. G. *J. Comput. Chem.* **1987**, *8*, 462.
(22) Mezey, P. G. *Int. J. Quantum Quant. Biol. Symp.*, **1987**, *14*, 127.
(23) Mezey, P. G. *J. Math. Chem.* **1988**, *2*, 299.
(24) Walker, P. D.; Arteca, G. A.; Mezey, P. G. *J. Comput. Chem.* **1991**, *12*, 220.
(25) Walker, P. D.; Arteca, G. A.; Mezey, P. G. *J. Comput. Chem.*, in press.
(26) Mezey, P. G. *J. Math. Chem.* **1992**, *11*, 27.
(27) Mezey, P. G. *J. Math. Chem.* **1991**, *7*, 39.
(28) Mezey, P. G. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 650.
(29) Mezey, P. G., Ed. *Mathematical Modeling in Chemistry*; VCH Publishers: New York, 1991.
(30) Mezey, P. G. *Shape in Chemistry: An Introduction to Molecular Shape and Topology*; VCH Publishers: New York, 1993.
(31) Walker, P. D.; Mezey, P. G. *J. Chem. Inf. Comput. Sci.*, to be published.