

Shape-similarity relations based on topological resolution

Paul G. Mezey

*Institute for Advanced Study, Collegium Budapest, 1014 Budapest, Szentháromság u. 2, Hungary and
Mathematical Chemistry Research Unit, Department of Chemistry and Department of Mathematics and
Statistics, University of Saskatchewan, 110 Science Place, Saskatoon, SK, Canada, S7N 5C9
E-mail: mezey@sask.usask.ca*

Received 28 April 1999

Shape-similarities of electron density clouds of molecules provide important clues concerning chemical and physical properties, including information about their reactivities in biochemical systems. The concept of topological resolution is used for quantifying molecular similarities: within a hierarchy of finer and cruder topologies, the crudest topology that already provides discrimination between two objects (such as two fuzzy electron density clouds) is used to define a measure of their similarities. The finer this topology, the more similar the two objects. This approach, the method of topological resolution-based similarity measures (TRBSM), can be combined with a geometrically motivated resolution-based similarity measure (RBSM) within a metric space. Some of the relations between these two approaches are discussed in this contribution, with special emphasis on applications to electron densities.

1. Introduction

Early studies of the three-dimensional shapes of molecules were carried out within the context of stereochemistry (see, e.g., [1,3–5,8,12,26,28–30] where the emphasis was placed on the nuclear arrangements. However, chemists have realized early that it is not the nuclei but the electron density cloud that is the most chemically relevant component of molecules. Nevertheless, the shift of attention to the electronic charge cloud has occurred only slowly, since the study of low electron densities within the bonding regions of molecules has proved to be a difficult task, both experimentally and theoretically.

The mathematical tools that appear eminently suitable for the theoretical and computational analysis of electron density clouds are provided by topology (for a sample of the relevant branches of topology, see, e.g., [2,6,7,9,11,27,31–34]). Within the topological context one should note that the physical nature of molecular electron densities provides special opportunities to exploit both topological and more conventional geometrical methods for shape analysis. In this contribution such an approach will be followed.

In recent studies the information content of various local and global representations of molecular electron density clouds were the subject of special scrutiny, initiated

by the proof of the Holographic Electron Density Theorem [22,23]. This theorem states that any nonzero volume piece of the (nondegenerate, ground state) electron density cloud of any molecule contains the complete information about the boundaryless electron density of the entire molecule. This statement can be viewed as an actual strengthening of the celebrated Hohenberg–Kohn theorem [10], where the latter states that the nondegenerate ground state electron density of a molecule determines the energy and other properties of the molecule. Based on the Holographic Electron Density Theorem, it is now evident that any small nonzero volume piece of the electron density already determines the energy and other properties of the complete molecule. That is, there is no need for the complete electron density (in the sense of the original Hohenberg–Kohn theorem) to determine these other properties. Various applications of this new theorem can already be found in the literature [24,25].

In the present study the information content of molecular electron densities will be viewed from the perspective of topological shape analysis. Topology is a powerful mathematical tool for shape characterization of fuzzy electron density clouds [18–21], a fact that serves as the basis of the molecular Shape Group approaches [13–17], relying on the homology groups of algebraic topology. In fact, the shape group methods convert some of the essential aspects of molecular shape information into a series of topological invariants, such as the ranks of homology groups defined by the interplay between the local and global curvature properties of the electron density. In this conversion, a formal reduction of information does occur, however, this can be controlled by the choice of the range and representation of parameters describing electron density and local curvature information.

In the context of electron density, the level sets $G(K, a)$ (often referred to as molecular isodensity contours, MIDCOs)

$$G(K, a) = \{\mathbf{r}: \rho(K, \mathbf{r}) = a\} \quad (1)$$

are of special importance, where the density threshold a can take values from the $[0, \infty)$ interval. (Note that in practice only a finite interval $[a_{\min}, a_{\max}]$ is considered.)

For a simple shape analysis, the local shape of each MIDCO $G(K, a)$ is compared to a range of reference curvatures b . Accordingly, all points \mathbf{r} along each MIDCO $G(K, a)$ are classified according to these local curvatures b . Specifically, there are three domain types, $D_2(b)$, $D_1(b)$ and $D_0(b)$. At point \mathbf{r} the contour surface $G(K, a)$ belongs either to

- (i) a domain of type $D_2(b)$, if at point \mathbf{r} the MIDCO $G(K, a)$ is locally convex relative to reference curvature b , or to
- (ii) a domain of type $D_1(b)$, if at point \mathbf{r} the MIDCO $G(K, a)$ is locally of the saddle type relative to reference curvature b , or to
- (iii) a domain of type $D_0(b)$, if at point \mathbf{r} the MIDCO $G(K, a)$ is locally concave relative to reference curvature b .

The patterns generated by these domains are analyzed and characterized using the tools of topology. In this approach, the concepts of cruder–finer topologies are

of relevance, defined in terms of conveniently chosen subbases, where the following conventions and notations are used.

We say that within a set X a topology \mathbf{T} is defined if a family of subsets of X is specified as the open sets in X , where these sets must fulfill the following mutual compatibility conditions.

A family \mathbf{T} of subsets of X ,

$$\mathbf{T} = \{T_\alpha: X \supset T_\alpha\}, \quad (2)$$

is called a topology on set X , if

$$(i) \quad X, \emptyset \in \mathbf{T}, \quad (3)$$

where \emptyset is the empty set,

$$(ii) \quad \bigcup_{\beta} T_\beta \in \mathbf{T} \quad (4)$$

for any number of sets in \mathbf{T} , and

$$(iii) \quad T_\alpha \cap T_\beta \in \mathbf{T} \quad (5)$$

for any two sets $T_\alpha, T_\beta \in \mathbf{T}$.

A base \mathbf{B} of topology \mathbf{T} is a family of subsets of X where any element

$$T_\alpha \in \mathbf{T} \quad (6)$$

of \mathbf{T} can be obtained as the union of sets from base \mathbf{B} :

$$T_\alpha = \bigcup_{\gamma} B_\gamma \in \mathbf{B}. \quad (7)$$

A subbase \mathbf{S} of topology \mathbf{T} is a family of subsets of X where any element

$$B_\gamma \in \mathbf{B} \quad (8)$$

of a base \mathbf{B} of topology \mathbf{T} can be obtained as finite intersection of sets from subbase \mathbf{S} :

$$B_\gamma = \bigcap_{\delta} S_\delta, \quad \text{for finitely many } S_\delta \in \mathbf{S}. \quad (9)$$

Clearly, the level within a hierarchy of finer–cruder topologies, as well as the level of topological resolution can be controlled by the choices of subbases \mathbf{S} and bases \mathbf{B} , leading to a series of different topologies \mathbf{T} defined on the same underlying space X . In particular, if for two generating subbases \mathbf{S}_1 and \mathbf{S}_2 containing families of subsets from the same underlying space X the relation

$$\mathbf{S}_2 \supset \mathbf{S}_1 \quad (10)$$

holds, then for the corresponding topologies generated by these subbases the relation

$$\mathbf{T}_2 \supset \mathbf{T}_1 \quad (11)$$

must hold, that is, topology \mathbf{T}_2 is finer than topology \mathbf{T}_1 .

We shall consider a family \mathbf{T} of topologies \mathbf{T}_i given in the underlying set X ,

$$\mathbf{T} = \{\mathbf{T}_1, \mathbf{T}_2, \dots, \mathbf{T}_i, \mathbf{T}_{i+1}, \dots\}, \quad (12)$$

and we further assume that these topologies \mathbf{T}_i are fully ordered by the finer–cruder relation, that is,

$$\mathbf{T}_{i+1} \supset \mathbf{T}_i \quad (13)$$

for every index-pair i and $i+1$ represented in family \mathbf{T} . The above finer–cruder relation can be expressed for the corresponding topological spaces (X, \mathbf{T}_{i+1}) and (X, \mathbf{T}_i) as

$$(X, \mathbf{T}_{i+1}) \supset (X, \mathbf{T}_i) \quad (14)$$

for every index-pair i and $i+1$, for which \mathbf{T}_i and \mathbf{T}_{i+1} are included in the family \mathbf{T} of topologies.

Our task is to use the series of topological spaces

$$\dots \supset (X, \mathbf{T}_{i+1}) \supset (X, \mathbf{T}_i) \supset \dots \supset (X, \mathbf{T}_2) \supset (X, \mathbf{T}_1) \quad (15)$$

for the description of similarities of objects defined within the underlying space X , and the finer–cruder relation in series (15) provides the tools for this purpose.

2. The construction of subbases for hierarchies of topologies

The $D_2(b)$, $D_1(b)$, and $D_0(b)$ domains of MIDCOs (collectively referred to as the D_μ domains) serve as natural choices for defining a series of topologies. For a given choice of the curvature parameter b (by itself a geometrical entity), the $D_2(b)$, $D_1(b)$, and $D_0(b)$ domains of a MIDCO $G(K, a)$ form a family of subsets of the MIDCO, and this family

$$\mathbf{S}(K, a, b) = \{D_\mu(b)\} \quad (16)$$

may be taken as a subbase for a topology $\mathbf{T}(K, a, b)$ on $G(K, a)$.

Take a finite series of monotonically increasing b values,

$$b_1 < b_2 < \dots < b_i < \dots < b_m, \quad (17)$$

and the corresponding series of the associated truncation patterns

$$P(K, a, b_1), P(K, a, b_2), \dots, P(K, a, b_i), \dots, P(K, a, b_m) \quad (18)$$

on the MIDCO $G(K, a)$.

One may take the combined pattern of several of these patterns, for example, by superimposing all the first i' patterns, i.e. the subsequence of patterns $P(K, a, b_i)$ for indices from 1 to i' :

$$P(K, a, b_1), P(K, a, b_2), \dots, P(K, a, b_{i'}). \quad (19)$$

Using these patterns, the combined pattern

$$P(K, a, b_1, \dots, b_{i'}) \quad (20)$$

of the MIDCO $G(K, a)$ is obtained.

Just as for each b_i value the set of $D_\mu(b_i)$ domains of the pattern $P(K, a, b_i)$ may serve as a defining subbase $\mathbf{S}(K, a, b_i)$

$$\mathbf{S}(K, a, b_i) = \{D_\mu(b_i)\}, \quad (21)$$

leading to a specific topology $\mathbf{T}(K, a, b_i)$ on the MIDCO $G(K, a)$, also, for each series $b_1 < b_2 < \dots < b_i$, the union of subbases $\mathbf{S}(K, a, b_1)$, $\mathbf{S}(K, a, b_2)$, \dots , $\mathbf{S}(K, a, b_{i'})$,

$$\mathbf{S}(K, a, b_1, \dots, b_{i'}) = \bigcup_{k=1}^{i'} \mathbf{S}(K, a, b_k), \quad (22)$$

itself can be regarded as a subbase that defines a topology $\mathbf{T}(K, a, b_1, \dots, b_{i'})$, where the corresponding topological spaces are

$$(G(K, a), \mathbf{T}(K, a, b_i)) \quad (23)$$

and

$$(G(K, a), \mathbf{T}(K, a, b_1, \dots, b_{i'})), \quad (24)$$

respectively.

This construction of topologies on the MIDCO $G(K, a)$ ensures that for any increasing series of $b_{i'}$ values the corresponding topologies are related by the weaker–stronger relation.

Specifically, for the defining subbases $\mathbf{S}(K, a, b_1, \dots, b_{i'})$ of the topologies $\mathbf{T}(K, a, b_1, \dots, b_{i'})$ of various upper indices i' fulfilling the relation

$$1 \leq i' \leq m - 1, \quad (25)$$

the following inclusion relation:

$$\mathbf{S}(K, a, b_1, \dots, b_{i'+1}) \supset \mathbf{S}(K, a, b_1, \dots, b_{i'}), \quad (26)$$

holds. Consequently, the corresponding topologies defined by these subbases also obey an inclusion relation (stronger–weaker relation)

$$\mathbf{T}(K, a, b_1, \dots, b_{i'+1}) \supset \mathbf{T}(K, a, b_1, \dots, b_{i'}). \quad (27)$$

In other words, the series of subbases $\mathbf{S}(K, a, b_1, \dots, b_{i'})$, the corresponding topologies $\mathbf{T}(K, a, b_1, \dots, b_{i'})$, and the associated topological spaces $(G(K, a), \mathbf{T}(K, a, b_1, \dots, b_{i'}))$ are fully ordered,

$$\begin{aligned} \mathbf{S}(K, a, b_1, \dots, b_m) \supset \mathbf{S}(K, a, b_1, \dots, b_{m-1}) \supset \dots \\ \supset \mathbf{S}(K, a, b_1, \dots, b_{i'}) \supset \dots \supset \mathbf{S}(K, a, b_1), \end{aligned} \quad (28)$$

$$\begin{aligned} \mathbf{T}(K, a, b_1, \dots, b_m) \supset \mathbf{T}(K, a, b_1, \dots, b_{m-1}) \supset \dots \\ \supset \mathbf{T}(K, a, b_1, \dots, b_{i'}) \supset \dots \supset \mathbf{T}(K, a, b_1), \end{aligned} \quad (29)$$

and

$$\begin{aligned} (G(K, a), \mathbf{T}(K, a, b_1, \dots, b_m)) \supset (G(K, a), \mathbf{T}(K, a, b_1, \dots, b_{m-1})) \supset \dots \\ \supset (G(K, a), \mathbf{T}(K, a, b_1, \dots, b_{i'})) \supset \dots \supset (G(K, a), \mathbf{T}(K, a, b_1)). \end{aligned} \quad (30)$$

The above ordering, by its very construction in terms of subbases ordered by the inclusion relations (28), represents a hierarchy of finer–cruder topologies.

3. Topological resolution as a similarity measure

The monotonic series (29) of topologies provides a convenient basis for an implementation of the technique of topological resolution for the shape analysis and eventual similarity analysis of MIDCOs $G(K, a)$.

The index i' serves as a monotonically changing integer parameter, leading gradually to increasing topological resolutions. Accordingly, a similarity measure based on topological resolutions can be constructed using these topologies $\mathbf{T}(K, a, b_1, \dots, b_{i'})$. One can recognize that this similarity measure is a special case of resolution-based similarity measures, RBSM. When it is necessary to emphasize that this is a *topological* resolution-based similarity measure, then the acronym TRBSM may be used.

Consider two molecules M_1 and M_2 , in two nuclear configurations, K_1 and K_2 , respectively. Restrict the analysis to two density thresholds a_1 and a_2 , and to the corresponding MIDCOs $G(K_1, a_1)$ and $G(K_2, a_2)$, respectively. The associated topological resolution-based similarity measure is constructed as follows.

We assume that for the range $1 \leq i' \leq m$ of indices i' , two series of topologies,

$$\{\mathbf{T}(K_1, a_1, b_1, \dots, b_{i'})\} \quad (31)$$

and

$$\{\mathbf{T}(K_2, a_2, b_1, \dots, b_{i'})\}, \quad (32)$$

are determined for the molecules M_1 and M_2 , respectively.

Within the context of the series of topologies discussed above, we say that two MIDCOs $G(K_1, a_1)$ and $G(K_2, a_2)$ have $[b_1, \dots, b_{i'}]$ -equivalent shapes at some level i' of topological resolution,

$$G(K_1, a_1) \text{ eq}[b_1, \dots, b_{i'}] G(K_2, a_2), \quad (33)$$

if and only if there is a one to one and onto correspondence between the two defining subbases $\mathbf{S}(K_1, a_1, b_1, \dots, b_{i'})$ and $\mathbf{S}(K_2, a_2, b_1, \dots, b_{i'})$ where the curvature index μ assignment of each element of the subbase for each sublevel k , $1 \leq k \leq i'$ is preserved.

If the curvature values are evident from the context, then one may use the simpler notation

$$G(K_1, a_1) \text{ eq } G(K_2, a_2). \quad (34)$$

Molecular electron densities show many common features. If the first reference curvature b_1 is chosen as a large enough negative value, and if the electron density thresholds a_1 and a_2 of the two MIDCOs are chosen appropriately, then the equivalence

$$G(K_1, a_1) \text{ eq}[b_1] G(K_2, a_2) \quad (35)$$

holds for any two nuclear configurations K_1 and K_2 of practically any two molecules M_1 and M_2 . This is a consequence of the simple fact that for small enough density thresholds a_1 and a_2 , both of the MIDCOs $G(K_1, a_1)$ and $G(K_2, a_2)$ are, typically, topological spheres. This occurs if both patterns $P(K_1, a_1, b_1)$ and $P(K_2, a_2, b_1)$ are trivial patterns, each formed by a single domain $D_0(b_1)$. Consequently, a common initial pattern exists for any two molecules M_1 and M_2 , if taken at a low level of topological resolution. At this low resolution level, the two molecular shapes appear topologically equivalent, not dependent either on the chemical nature or on the actual nuclear configurations K_1 and K_2 .

However, if the two molecules M_1 and M_2 are different, or if one considers two different nuclear configurations K_1 and K_2 for a given molecule, then not all levels of combined patterns are topologically equivalent. Consequently, the method of topological resolution reveals the differences between the two molecules, unless the interval of the curvature threshold values is not sufficiently resolved by the selection $b_1 < b_2 < \dots < b_i < \dots < b_m$.

Typically, a gradual increase of the topological resolution eventually reveals the shape differences between two different molecules M_1 and M_2 . If the reference curvature $b_{i'}$ has a high enough value, then even slight conformational deviations of the same molecule are revealed by the method of topological resolution.

A non-equivalence of the shapes of two MIDCOs $G(K_1, a_1)$ and $G(K_2, a_2)$ at some level i' of topological resolution is formally stated as

$$G(K_1, a_1) \text{ noneq } [b_1, \dots, b_{i'}] G(K_2, a_2). \quad (36)$$

If the *lowest* index i' for an equivalence $G(K_1, a_1) \text{ eq}[b_1, \dots, b_{i'}] G(K_2, a_2)$ is higher, this implies a higher level of similarity, where a finer (stronger) topology provides a higher level of topological resolution.

Consider a family of molecules

$$M_1, M_2, \dots, M_j, \dots, M_n, \quad (37)$$

with nuclear configurations

$$K_1, K_2, \dots, K_j, \dots, K_n, \quad (38)$$

and (the comparable) electron density threshold values

$$a_1, a_2, \dots, a_j, \dots, a_n, \quad (39)$$

for a corresponding set of MIDCOs

$$G(K_1, a_1), G(K_2, a_2), \dots, G(K_j, a_j), \dots, G(K_n, a_n). \quad (40)$$

For the pairwise comparisons of these MIDCOs, the maximum indices of eq[$b_1, \dots, b_{i'}$]-equivalence are denoted by

$$i_{jk} = i_{jk} [G(K_j, a_j), G(K_k, a_k)]. \quad (41)$$

If, for example, for the three indices

$$i_{12} < i_{13} < i_{23} \quad (42)$$

holds, then among the three MIDCO surfaces $G(K_1, a_1)$, $G(K_2, a_2)$, and $G(K_3, a_3)$ of three molecules, M_1 , M_2 , and M_3 , taken at nuclear arrangements K_1 , K_2 , and K_3 , and using the method of topological resolution, the pair $G(K_2, a_2)$ and $G(K_3, a_3)$ of MIDCOs shows the highest degree of similarity, and the method of topological resolution indicates the lowest degree of similarity for the pair $G(K_1, a_1)$ and $G(K_2, a_2)$ of MIDCOs.

4. A topological dissimilarity measure based on topological resolution

For a fixed set of topologies, the quantity

$$m - i_{12}$$

may serve as a dissimilarity measure. This quantity, however, is not a metric on the abstract space of all shapes, since a formal zero distance may be obtained for two geometrically different objects, that is, for two objects of different shapes at infinite resolution.

5. Summary

An implementation of the framework of topological resolution, adapted to the description of molecular similarity and dissimilarity, is described.

Acknowledgements

This study was supported by funds from the Institute for Advanced Study, Collegium Budapest, Hungary, and by a research grant from the Natural Sciences and Engineering Research Council of Canada.

References

- [1] *Topics in Stereochemistry*, Vol. 1, eds. N.L. Allinger and E.L. Eliel (Wiley, New York, 1967).

- [2] R.L. Bishop and R.J. Crittenden, *Geometry of Manifolds* (Academic Press, New York, 1964).
- [3] E.L. Eliel, *Stereochemistry of Carbon Compounds* (McGraw-Hill, New York, 1962).
- [4] E.L. Eliel, *Elements of Stereochemistry* (Wiley, New York, 1969).
- [5] K. Freudenberg, *Stereochemie* (F. Deuticke, Leipzig, Wien, 1932).
- [6] T.W. Gamelin and R.E. Greene, *Introduction to Topology* (Saunders College Publishing, New York, 1963).
- [7] M. Greenberg, *Lectures on Algebraic Topology* (Benjamin, New York, 1967).
- [8] J. Grundy, *Stereochemistry: The Static Principles* (Butterworths, London, 1964).
- [9] V. Guillemin and A. Pollack, *Differential Topology* (Prentice Hall, Englewood Cliffs, 1974).
- [10] P. Hohenberg and W. Kohn, Inhomogeneous electron gas, *Phys. Rev.* 136 (1964) B864–B871.
- [11] S.-T. Hu, *Elements of General Topology* (Holden-Day, San Francisco, 1969).
- [12] W. Klyne, *Progress in Stereochemistry* (Academic, New York, 1954).
- [13] P.G. Mezey, Group theory of electrostatic potentials: A tool for quantum chemical drug design, *Int. J. Quantum Chem. Quant. Biol. Symp.* 12 (1986) 113.
- [14] P.G. Mezey, The shape of molecular charge distributions: Group theory without symmetry, *J. Comput. Chem.* 8 (1987) 462.
- [15] P.G. Mezey, Group theory of shapes of asymmetric biomolecules, *Int. J. Quantum Chem., Quant. Biol. Symp.* 14 (1987) 127.
- [16] P.G. Mezey, Global and local relative convexity and oriented relative convexity; application to molecular shapes in external fields, *J. Math. Chem.* 2 (1988) 325.
- [17] P.G. Mezey, Shape group studies of molecular similarity: Shape groups and shape graphs of molecular contour surfaces, *J. Math. Chem.* 2 (1988) 299.
- [18] P.G. Mezey, The degree of similarity of three-dimensional bodies; applications to molecular shapes, *J. Math. Chem.* 7 (1991) 39–49.
- [19] P.G. Mezey, *Shape in Chemistry: An Introduction to Molecular Shape and Topology*, (VCH Publishers, New York, 1993).
- [20] P.G. Mezey, Quantum chemistry of macromolecular shape, *Internat. Rev. Phys. Chem.* 16 (1997) 361–388.
- [21] P.G. Mezey, Shape in quantum chemistry, in: *Conceptual Trends in Quantum Chemistry*, Vol. 3, eds. J.-L. Calais and E.S. Kryachko (Kluwer, Dordrecht, 1997) pp. 519–550.
- [22] P.G. Mezey, Generalized chirality and symmetry deficiency, *J. Math. Chem.* 23 (1998) 65–84.
- [23] P.G. Mezey, The Holographic Electron Density Theorem and quantum similarity measures, *Mol. Phys.* 96 (1999) 169–178.
- [24] P.G. Mezey, Holographic Electron Density Shape Theorem and its role in drug design and toxicological risk assessment, *J. Chem. Inf. Comp. Sci.* 39 (1999) 224–230.
- [25] P.G. Mezey, R. Ponec, L. Amat and R. Carbo-Dorca, Quantum similarity approach to the characterization of molecular chirality, *Enantiomers* 4 (1999) 371–378.
- [26] K. Mislow, *Introduction to Stereochemistry* (Benjamin, New York, 1966).
- [27] M. Morse and S.S. Cairns, *Critical Point Theory in Global Analysis and Differential Topology: An Introduction* (Academic Press, New York, London, 1969).
- [28] M.S. Newman, *Steric Effects in Organic Chemistry* (Wiley, New York, 1956).
- [29] M. Nógrádi, *Stereochemistry: Basic Concepts and Applications* (Pergamon Press, Oxford, 1981).
- [30] J. Rétey and J.A. Robinson, *Stereospecificity in Organic Chemistry and Enzymology* (Verlag Chemie, Weinheim, 1982).
- [31] G.F. Simmons, *Introduction to Topology and Modern Analysis* (McGraw-Hill, New York, 1963).
- [32] I.M. Singer and J.A. Thorpe, *Lecture Notes on Elementary Topology and Geometry* (Springer-Verlag, New York, 1976).
- [33] E.H. Spanier, *Algebraic Topology* (McGraw-Hill, New York, 1966).
- [34] J. Vick, *Homology Theory* (Academic Press, New York, 1973).