STUDIA UNIVERSITATIS BABEŞ-BOLYAI, CHEMIA, LIII, 2, 2008

MEASURES OF SHAPE SIMILARITY OF ELECTRON DENSITY CLOUDS IN MOLECULAR MODELING

PAUL G. MEZEY[a,b,*], CORNELIA MAJDIK[c]

ABSTRACT. Shape similarity measures for molecular electron densities may serve many purposes, one important application of such measures is in QSAR, where precise electron density comparisons provide high levels of correlations with various biochemical activities. Some shape similarity measures are "strictly shape-based", in the sense that they ignore all size information, and they can be considered "pure" shape similarity measures. In this contribution it is argued that shape similarity measures which retain some size information are likely to be better suited for some molecular applications, as a consequence of the relatively narrow range of the typical bond lengths, the typical sizes of functional groups, and the limited, but non-negligible local size variations of the associated electron density clouds due to the influence of their local surroundings.

INTRODUCTION

Similarity in chemistry plays a fundamental role; one may claim that the original understanding of most laws of chemistry has been based on the recognition of similarities between seemingly different molecules or phenomena. In searching for the explanation of those similarities, new theories and new laws have been discovered, hence similarity is fundamental in the development of chemistry, and in general, all sciences.

In recent years there has been a considerable amount of scientific work devoted to molecular similarity; for some of the main directions the reader may consult references [1-10]. In particular, QSAR approaches and drug design [11-14] have provided strong motivation to place similarity analysis on a strong foundation.

[a,*] *Scientific Modeling and Simulation Laboratory (SMSL), Department of Chemistry and Department of Physics and Physical Oceanography, Memorial University of Newfoundland, 283 Prince Philip Drive, St. John's, NL A1B 3X7, Canada, paul.mezey@gmail.com*

[b] *Institute for Advanced Study, Collegium Budapest, Szentháromság utca 2, 1014 Budapest, Hungary*

[c] *Babes-Bolyai University, Faculty of Chemistry and Chemical Engineering, Cluj-Napoca, 1, Kogalniceanu, RO-400084, Romania*

We shall focus on a particular, intuitively rather straightforward shape similarity measure, the so-called Scaling-Nesting Similarity Measure, SNSM, where the separation of shape and size features are especially clear [15,16]; this similarity measure has also provided motivation for some of the alternative, quantum chemical approaches [17].

SCALING-NESTING SIMILARITY MEASURES WITH AND WITHOUT PRE-SCALING

The main idea of the scaling-nesting similarity measure is very simple. If one restricts the analysis to objects of the same size, for example, to objects of the same volume or to planar objects of the same area, then if two objects A and B have exactly the same shape, then A fits within B and B fits within A perfectly. If they have different shapes, then each objects must be reduced in size, in order to fit within the other, and the maximal scaling factor that already allows one to fit within the other provides a numerical characterization of the difference in their shapes. This idea is illustrated by the two planar objects A and B in Figure 1.

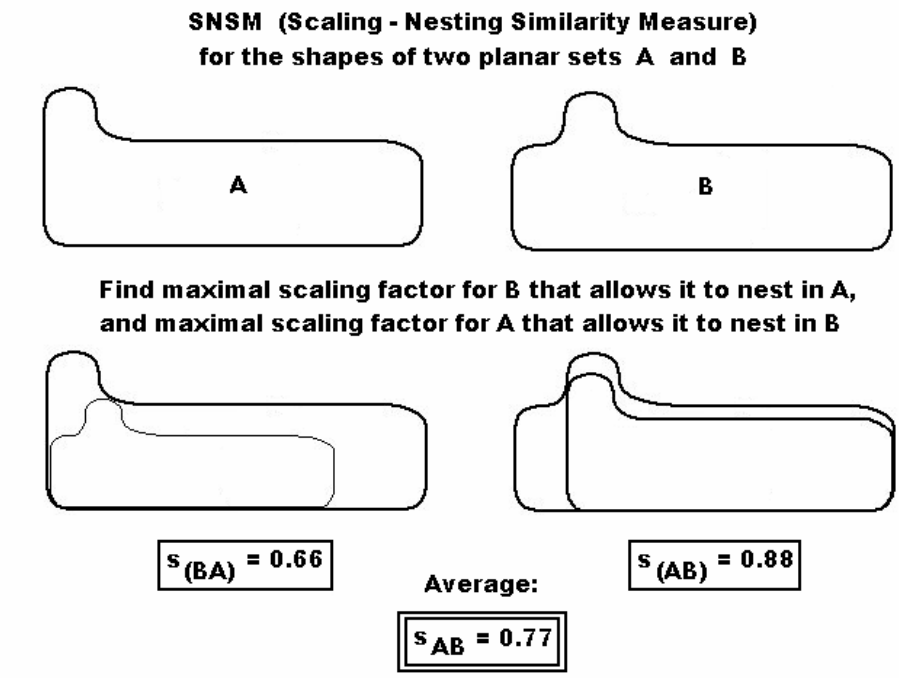

Figure 1. The Scaling-Nesting Similarity Measure applied to two planar objects A and B of the same area (same "size")

Usually, the two objects A and B require different scaling factors, hence a similarity measure based on such scaling initially provides only an asymmetric similarity measure, also called “Semi-Similarity Measure” [15]. Such asymmetric similarity measures may serve a useful purpose, when similarity may be interpreted differently from A to B than from B to A, for example, if similarity between languages of two populations of different mother tongues are analyzed in terms of the percentage of one population understanding the language of the other population, these percentages are often different, as it is the case for Spanish and Portuguese. However, in most cases one may prefer a measure that is symmetric, providing the same similarity value independently of the ordering of the objects of study. The simplest such measure is the average of the two numerical values of such asymmetric measures; evidently, the average must always be symmetric.

We denote the maximum scaling factor that allows A to fit within B by $s_{(A,B)}$, similarly, the maximum scaling factor that allows B to fit within A is denoted by $s_{(B,A)}$, and the average of the two factors by s_{AB},

$$s_{AB} = (s_{(A,B)} + s_{(B,A)}) / 2 \quad (1)$$

a function evidently symmetric for the interchange of A and B.

We can often formulate property correlations easier in terms if dissimilarity than in terms of similarity. The associated Scaling-Nesting Dissimilarity Measure can be defined as

$$S_{AB} = 1 - s_{AB} = 1 - (s_{(A,B)} + s_{(B,A))} / 2 \quad (2)$$

Evidently, this Scaling-Nesting Dissimilarity Measure S_{AB} is also symmetric for the interchange of A and B.

For this Scaling-Nesting Dissimilarity Measure, it is easy to prove an even stronger statement, not restricted to symmetry. In fact, the Scaling-Nesting Dissimilarity Measure is a proper metric, fulfilling all the general mathematical conditions for a “distance-like” function:

$S_{AB} \geq 0$,	(non-negativity)	(3)
$S_{AB} = 0$ if and only if A=B	(identity)	(4)
$S_{AB} = S_{BA}$	(symmetry)	(5)
$S_{AC} \leq S_{AB} + S_{AB}$	(triangle inequality)	(6)

The first three conditions are trivially fulfilled, and the triangle inequality can also be verified [16]; a simple demonstration is given below.

The triangle inequality we want to prove, inequality (6), can be written explicitly in terms of the symmetric similarity measures as

$$1 - s_{AC} \leq 1 - s_{AB} + 1 - s_{BC} \quad (7)$$

and in terms of the semi-similarity measures as

$$1 - (s_{(A,C)} + s_{(C,A))} / 2 \quad \leq 1 - (s_{(A,B)} + s_{(B,A))} / 2 + 1 - (s_{(B,C)} + s_{(C,B)} / 2 \qquad (8)$$

that reduces to

$$s_{(A,B)} + s_{(B,C)} - s_{(A,C)} + s_{(C,B)} + s_{(B,A)} - s_{(C,A)} \leq 2 \; . \qquad (9)$$

In order to prove inequality (9), let us first consider two ways of scaling A to fit within C: directly, with maximum scaling factor $s_{(A,C)}$, or in two steps, first scaling A to fit within B, then scaling B (with the scaled A included) to fit within C. Evidently, this second procedure will also result in A being scaled so it is included in C, however, this second process is less efficient, and the overall scaling factor for this process is the product of the two scaling factors, $s_{(A,B)}\, s_{(B,C)}$, where this product cannot exceed $s_{(A,C)}$, this latter being the maximum scaling factor. Consequently,

$$s_{(A,B)}\, s_{(B,C)} \leq s_{(A,C)} \qquad (10)$$

We also realize that for the positive quantities bounded by 1,

$$0 \leq s_{(A,B)} \, , \; s_{(B,C)} \leq 1 \qquad (11)$$

the inequality

$$0 \leq (s_{(A,B)} - 1)\,(s_{(B,C)} - 1) \qquad (12)$$

must always hold, that gives

$$s_{(A,B)} + s_{(B,C)} - s_{(A,B)}\, s_{(B,C)} \leq 1 \qquad (13)$$

Due to inequality (10), the product $s_{(A,B)}\, s_{(B,C)}$ cannot be greater than $s_{(A,C)}$, consequently, replacing the former with the latter in inequality (13) only strengthens this inequality, and

$$s_{(A,B)} + s_{(B,C)} - s_{(A,C)} \leq 1 \; . \qquad (14)$$

Precisely the same way we obtain

$$s_{(C,B)} + s_{(B,A)} - s_{(C,A)} \leq 1 \; . \qquad (15)$$

The sum of the two inequalities (14) and (15) then gives inequality (9), that proves the triangle inequality for the Scaling Nesting Dissimilarity Measure, and also completes the verification that it is indeed a metric.

This dissimilarity measure S_{AB} is a “pure” shape measure, as it is independent of size; in fact, for its application to objects of different size, a pre-scaling to a common size is needed, otherwise the scaling factors may take any values, independent of the actual shapes involved.

Whereas skipping pre-scaling to a common size may render some of the advantageous properties of the metric invalid, nevertheless, without pre-scaling the dissimilarity measure involves a natural size-dependence, that is also providing some advantages in the chemical context. One also may define an asymmetric dissimilarity measure directly, by turning the semi-similarity measures into semi-dissimilarity measures, before symmetrizing by using averages. For example, as illustrated in Figures 2 and 3, one may consider the dissimilarities of water, methanol, and ethanol in terms of molecular isodensity contours (MIDCO's), for example, using the a = 0.1 a.u. (atomic unit) isocontours. Reducing water to fit within ethanol provides a more chemically relevant measure than scaling ethanol to fit within water, yet due to the nearly equivalent sizes of the essential OH groups, a pre-scaling of these two molecules for the same size has very little chemical relevance. One may deduce similar conclusions when considering the other two pairs in this molecular family, water and methanol, or methanol and ethanol. Whereas the sizes of the chemically important common functional groups, in our case, the sizes of the OH groups are rather similar, by themselves not justifying pre-scaling, nevertheless, the sizes of the complete molecules are rather different, and pre-scaling for the complete sizes would cause major size differences in the OH groups, masking their high degree of chemical similarity. Without pre-scaling, one may always adopt the larger of the two scaling factors of the semi-similarity measures as being the chemically more relevant, for example, the scaling factor needed to fit water within ethanol, that would exclude the other scaling factor obtained by the drastic scaling ethanol to fit within water.

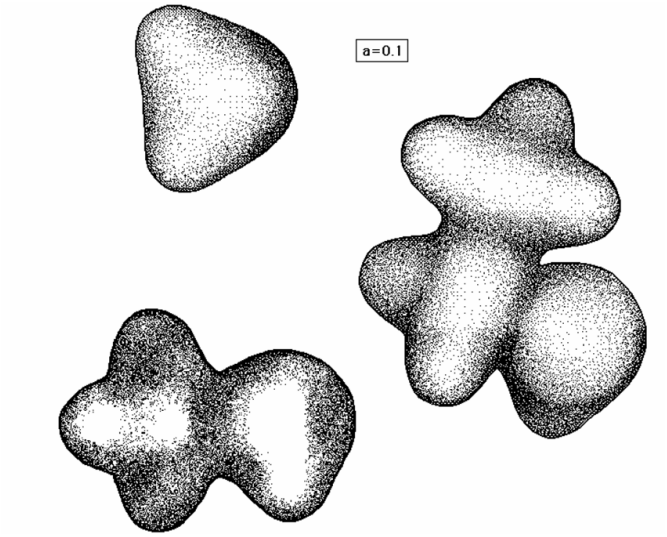

Figure 2. Electron density contours (MIDCO's) for water, methanol, and ethanol, 0.1 a.u. (atomic unit) contours.

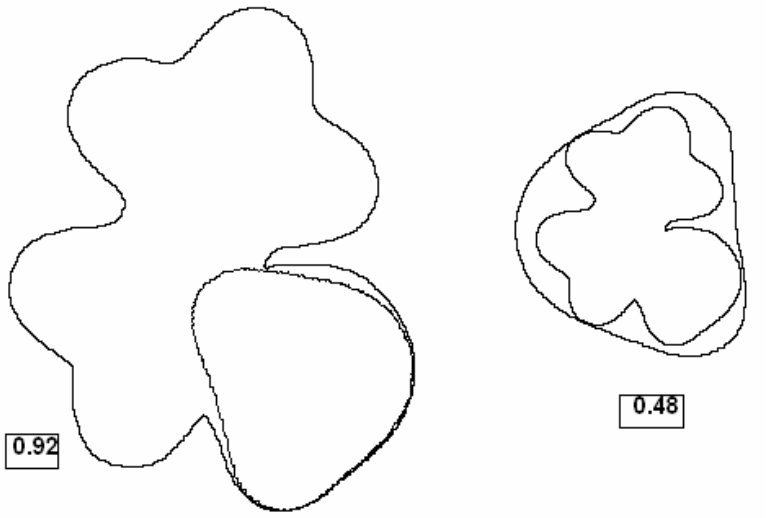

Figure 3. Without pre-scaling, some scaling-nesting relations for molecules of water and ethanol. For the water – ethanol pair, the two semi-similarity scaling factors are 0.92 and 0.48.

In conclusion, neither pre-scaling, nor the scaling of the larger molecule (or, if similarity is evaluated for local electron densities, the larger molecular fragment, or larger functional group) provide a chemically easily interpretable measure, but without pre-scaling, the semi-similarity measure giving the *larger scaling factor* provides the chemically most relevant measure of actual, chemical similarity.

CONCLUSION

It is demonstrated that in some chemical problems the use of semi-similarity measures of the scaling-nesting family of similarity measures, where the pre-scaling step for a common size is avoided, provides a chemically useful similarity measure.

REFERENCES

1. R. Carbó, L. Leyda, M. Arnau, *Int. J. Quantum Chem.* **1980**, *17*, 1185.
2. E. E. Hodgkin, W. G. J. Richards, *Chem. Soc., Chem. Commun.* **1986**, 1342.
3. M. A. Johnson, G. M. Maggiora, *Eds. Concepts and Applications of Molecular Similarity;* Wiley & Sons: New York, **1990**.
4. P. G. Mezey, *J. Chem. Inf: Comput. Sci.* **1992**, *32*, 650.
5. P. G. Mezey, *Shape in Chemistry: An Introduction to Molecular Shape and Topology;* VCH Publishers: New York, **1993**.

6. K. Sen, *Ed. Molecular Similarity, Topics in Current Chemistry*; Springer-Verlag: Heidelberg, **1995**; Vol 173.
7. P. M. Dean, *Ed. Molecular Similarity in Drug Design*; Chapman & Hall - Blackie Publishers: Glasgow, **1995**.
8. R. Carbó, *Ed. Molecular Similarity and Reactivity: From Quantum Chemical to Phenomenological Approaches*; Kluwer Academic Publishers: Dordrecht, The Netherlands, **1995**.
9. E. Besalú, R. Carbó, J. Mestres, M. Solà, *Top. Curr. Chem.,* **1995**, *173*, 31.
10. R. Carbó-Dorca, P. G. Mezey, *Eds. Advances in Molecular Similarity*; JAI Press: Greenwich, Connecticut, **1996**; Vol 1.
11. Y. C. Martin, *Quantitative Drug Design: A Critical Introduction*; Marcel Dekker: New York, **1978**.
12. W. G. Richards, *Quantum Pharmacology*; Butterworths: New York, **1983**.
13. R. Franke, *Theoretical Drug Design Methods*; Elsevier: Amsterdam, **1984**.
14. P. M. Dean, *Molecular Foundations of Drug-Receptor Interaction*; Cambridge University Press: Cambridge, **1987**.
15. P. G. Mezey, *Int. J. Quantum Chem.*, **1994**, *51*, 255.
16. P. G. Mezey, *Int. J. Quantum Chem.*, ***1997***, *63*, 105.
17. P. G. Mezey, *Int. J. Quantum Chem.*, ***1997***, *63*, 39.