A Formal Comparison between Molecular Quantum Similarity Measures and Indices

David Robert and Ramon Carbó-Dorca*

Institute of Computational Chemistry University of Girona, 17071 Girona, Catalonia, Spain

Received November 25, 1997

In this paper two methods are described to establish a formal comparison between different molecular quantum similarity measures (MQSM) and indices (MQSI) to detect redundancies in the information produced. The methods used are the Procrustes analysis and a proximity measure involving standarized similarity vectors. A small set of molecules, the fluoro- and chloro-substituted methanes, are used as an illustrative example, and conclusions obtained by direct comparison of the similarity matrixes are retrieved.

1. INTRODUCTION

The definition of a similarity measure between two molecules is a problem that is studied frequently in modern chemistry. A great number of measures and indices based on very different molecular aspects exist in the literature,[1] and all of them try to quantify the comparison between the structure of the elements of a molecular set, or some aspects of their topology. In most cases, one has to decide on which type of similarity measure or index to use from an extensive set. This choice can be a particularly difficult task if one does not know the advantages and drawbacks of each measure/index and the kind of information they produce. In addition, most of these measures and indices are related to each other, resulting in redundant information in the study of a molecular system. In this paper we present two methodologies for the comparison between the different similarity measures and indices, giving quantitative measures of the proximity of the information they produce and allowing elimination of redundancies.

This work is basically focused on the comparison between the different types of similarity measures and indices that constitute the basis of the Molecular Quantum Similarity Theory,[2] which are analyzed in their application to the concrete case of fluoro- and chlorosubstituted methanes.

Molecular Quantum Similarity Theory basic values, measures (MQSM) and indices (MQSI),[2] are general enough to present diverse structural forms. All these possible degrees of freedom can be summarized by two essential ones: the operator chosen to configure the MQSM, and the type of manipulation performed from the matrix elements of these MQSM or the so-called MQSI. The ultimate objective of the theory, the comparison between different objects of a quantum system, depends therefore on which MQSM or MQSI is chosen. Because all these forms yield the comparison between the same quantum objects, and assuming the relationships between the elements are unique, it can then be concluded that the existing degrees of freedom in the theory are nothing but different viewpoints of the same underlying situation. To evidence and to quantify in some manner the proximity relationship between the different configurations, two proximity measures between the different types of MQSM and MQSI will be described.

2. DEFINITIONS

2.1. Molecular Quantum Similarity Measures (MQSM). Quantum similarity constitutes a fundamental tool for ordering and classifying quantum systems using a quantum mechanical descriptor; that is, their density functions. In the field of quantum chemistry, this tool has been applied mainly to molecular systems. Recently, however, quantum similarity has been extended to other quantum systems of physical interest, such as atomic nuclei.[3]

An *n* particle system can be characterized by an *n*-order density function containing all the information one can extract. Because of the evident difficulties in modeling and manipulating these functions, we will restrict our calculations to first-order density functions, assuming the derived loss of information. Thus, if two first-order density functions for two systems A and B are known, $\rho_A(\boldsymbol{r}_1)$ and $\rho_B(\boldsymbol{r}_2)$, then a MQSM can be defined as the following integral:

$$Z_{AB} = \int_{D_1,D_2} \rho_A(\boldsymbol{r}_1)\, \Omega(\boldsymbol{r}_1,\boldsymbol{r}_2)\, \rho_B(\boldsymbol{r}_2)\, d\boldsymbol{r}_1\, d\boldsymbol{r}_2 \qquad (1)$$

where $\Omega(\boldsymbol{r}_1,\boldsymbol{r}_2)$ is a definite positive operator. This operator can be chosen arbitrarily, but the most usual are:

(a) $\Omega(\boldsymbol{r}_1,\boldsymbol{r}_2) = \delta(\boldsymbol{r}_1 - \boldsymbol{r}_2)$, a Dirac delta distribution; MQSM derived in this way are called *overlap-like* MQSM.[2]

(b) $\Omega(\boldsymbol{r}_1,\boldsymbol{r}_2) = |\boldsymbol{r}_1 - \boldsymbol{r}_2|^{-1}$, yielding the *Coulomb-like* MQSM.[2]

(c) $\Omega(\boldsymbol{r}_1,\boldsymbol{r}_2) = |\boldsymbol{r}_1 - \boldsymbol{r}_2|^{-2}$, yielding the *gravitational-like* MQSM.[4]

(d) $\Omega(\boldsymbol{r}_1,\boldsymbol{r}_2) = \rho_C(\boldsymbol{r})$, another molecular density function, which yields the *triple-density* MQSM.[5]

Other operators could be chosen, provided they are definite positive. The results from MQSM calculations can be expressed in a matrix form, where the element Z_{AB} corresponds to the comparison of molecule A with molecule B. Obviously, the MQSM matrix is symmetric.

2.2. Molecular Quantum Similarity Indices (MQSI). Starting with MQSM, one can define various possible mathematical manipulations, the so-called MQSI. All the possible descriptions of MQSI belong to one of two classes: C−class or D−class indices.[2a] A C−class index is referred to as a *correlation-like index*; that is, ranging from

S0095-2338(97)00105-4 CCC: $15.00
Published on Web 04/07/1998

0 (total dissimilarity) to 1 (complete similarity), depending on the similarity measure associated with the two molecules. D−Class indices are *distance-like indices*, and range from 0 (total similarity) to infinity (complete dissimilarity). Some of these indices are presented as follows:

D−Class Indices:

(a) The Generalized Dissimilarity Index is expressed as

$$D_{AB}(k,x) = [k(Z_{AA} + Z_{BB})/2 - xZ_{AB}]^{1/2} \qquad x \in [0,k] \tag{2}$$

When $k = x = 2$, eq 2 reduces to the Euclidean Distance Index:[2a]

$$D_{AB} = \sqrt{Z_{AA} + Z_{BB} - 2Z_{AB}} \tag{3}$$

C−Class Indices:

(a) The Cosine-like Similarity Index (also called Carbó Index)[2] is

$$C_{AB} = Z_{AB}[Z_{AA}Z_{BB}]^{-1/2} \tag{4}$$

This index can be interpreted as a generalized cosine of the angle between Hilbert space functions ρ_A and ρ_B.

(b) Two other indices can be extracted from the following generalized formula:[4]

$$V_{AB}(k,x) = (k - x)Z_{AB}D_{AB}^{-2}(k,x) \qquad k \in [0,2] \tag{5}$$

When $k = 2$ and $x = 0$, one can extract the Hodgkin−Richards Index:[6]

$$H_{AB} = 2Z_{AB}[Z_{AA} + Z_{BB}]^{-1} \tag{6}$$

When $k = 2$ and $x = 1$, one can extract the Tanimoto Index:[7]

$$T_{AB} = Z_{AB}[Z_{AA} + Z_{BB} - Z_{AB}]^{-1} \tag{7}$$

(c) The Petke Index[8] (continuous form) can be defined as:

$$P_{AB} = Z_{AB}D_{AB}^{-1}(\infty) \tag{8}$$

where $D_{AB}(\infty)$ is the *infinite order distance index*, which is defined as:

$$D_{AB}(\infty) = \max(Z_{AA},Z_{BB}) \tag{9}$$

There are other possible indices,[9−11] but they will not be discussed here.

The C−class indices can be transformed into a D−class form, for instance, by means of eq 10:

$$\delta_{ij} = \sqrt{1 - s_{ij}^2} \tag{10}$$

where s_{ij} is an element belonging to any C−class index. It can be demonstrated[12] that a transformation of this kind yields a Euclidean distance dissimilarity. It is also important to note that the construction of this transformation yields the transformed C−class indices ranging from 0 (complete similarity) to 1 (total dissimilarity), not to infinity. Other transformations between classes are possible, but they will not be studied here.[10]

All these indices are nothing but scalings or normalizations of the original MQSM; thus, they do not provide new information about the relationships between the compared molecules.

3. PROXIMITY MEASURES BETWEEN MQSM AND MQSI

The most usual and simplest way to compare the different MQSM or MQSI is by direct observation of the similarity matrixes. Once a pair of molecules is compared, one can look at the values given by the different MQSM or MQSI and try to extract conclusions. However, a proximity measure between the QMSM and QMSI can be defined to get a quantitative value of the resemblance between them. These measures will then make possible the statistical treatment to extract the contained information and to obtain graphical representations of the proximities. Closeness of two points representing two MQSM or MQSI indicates proximity between the information they produce or even a mathematical connection between them.

Two proximity measures are presented here: the Procrustes statistic and a similarity measure involving standarized similarity matrixes. These measures are not unique, in fact other proximity measures between MQSM and MQSI have already been defined.[11]

3.1. Procrustes Analysis. Procrustes analysis[13] is a multidimensional scaling technique that compares configurations derived, for instance, from different Euclidean distance definitions. For its application, the different $n \times n$ MQSM matrixes will need to be transformed into Euclidean distances through eq 3. Furthermore, MQSI will all be converted into D−class indices through the transformation in eq 10.

Now suppose that a configuration of n points in a q dimensional Euclidean space (obtained, for instance, using Metric Classical Scaling[13,14]), with coordinates given by the $n \times q$ matrix $\mathbf{X}$, needs to be optimally matched to another configuration of n points in a p ($p \geq q$) dimensional Euclidean space with coordinate matrix $\mathbf{Y}$. To simplify the study, throughout the present paper we will consider $p = q = 5$, the first five principal coordinates of the systems, each one accounting for >90% of the variation between the molecules studied. It is assumed that the rth point in the first configuration is in one-to-one correspondence with the rth point in the second configuration. Thus, the points in the $\mathbf{X}$ space will be dilated, translated, rotated, and reflected to new coordinates $\boldsymbol{x}'_r$, where:

$$\boldsymbol{x}'_r = \delta\, \mathbf{R}^T \boldsymbol{x}_r + \boldsymbol{t} \tag{11}$$

The matrix $\mathbf{R}$ is orthogonal giving a rigid rotation, vector $\boldsymbol{t}$ is a rigid translation vector, and δ is the dilation. The motions are sought that minimize the sum of the distances between points,

$$M^2 = \sum_{r=1}^{n} ||\boldsymbol{y}_r - \boldsymbol{x}'_r|| = \sum_{r=1}^{n} (\boldsymbol{y}_r - \delta\, \mathbf{R}^T \boldsymbol{x}_r - \boldsymbol{t})^T (\boldsymbol{y}_r - \delta\, \mathbf{R}^T \boldsymbol{x}_r - \boldsymbol{t}) \tag{12}$$

Assessing the matching of the two configurations can be done using the minimized value of M^2, conveniently scaled to

ensure the symmetry between configurations **X** and **Y**.[13a] This value is known as the *Procrustes statistic*:

$$M_0^2 = 1 - \{\mathrm{tr}(\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X})^{1/2}\}^2/\{\mathrm{tr}(\mathbf{X}^T\mathbf{X})\ \mathrm{tr}(\mathbf{Y}^T\mathbf{Y})\} \quad (13)$$

The Procrustes statistic has a D-class form, ranging from 0 to 1. Then, a symmetric matrix $\mathbf{M_0^2}$ can be constructed, where the element $[M_0^2]_{PQ}$ corresponds to the Procrustes statistic relative to the comparison between the configurations derived from MQSM (or MQSI) *P* and *Q*. This matrix has a dissimilarity form, so it can be studied with multivariate analysis techniques. In particular, it can be projected onto a bi- or tridimensional space using classical scaling, or analyzed straightforwardly with cluster analysis algorithms.[15] It is important to note that Procrustes analysis does not compare MQSM or MQSI, but the point configurations derived from them.

3.2. Standarized Matrix-Into-Vector Proximity Measure. Another proximity measure between different MQSM or MQSI can be described in the following way: MQSM or MQSI matrixes are expressed in a column vector form, using only the nonredundant components (i.e., the upper or lower triangle of the symmetric matrixes). These vectors can be standarized by subtracting the mean value of their components and dividing by the standard deviation.

Then a similarity measure between these *matrixes-into-vectors* structures can be defined as a simple scalar product:

$$S_{PQ} = \hat{\mathbf{X}}_P\,\hat{\mathbf{X}}_Q = \sum_{i=1}^{N} \hat{X}_i^{(P)}\hat{X}_i^{(Q)} \quad (14)$$

This similarity measure can be transformed into a Euclidean distance through eq 3 and studied by multivariate analysis techniques as a common dissimilarity matrix. Classical scaling[13,14] and cluster analysis[15] can also be applied to the aforementioned case.

4. ANALYSIS OF THE DISSIMILARITY MATRIXES

Once the dissimilarity matrixes between MQSM and MQSI are constructed, three techniques will be used to extract the information they contain.

Metric Classical Scaling (Principal Coordinate Analysis). This method[13,14] considers the objects as points and provides a projection of the set in a low-dimensional space. The method starts with a dissimilarity matrix between an objects set and its objective is to find the Euclidean distances $\{d_{ij}\}$ that best match the original dissimilarities $\{\delta_{ij}\}$. If the dissimilarities are precisely Euclidean distances (derived, for instance, with eq 3), then it is possible to find a configuration of points ensuring the equality between distances and dissimilarities.[16] Metric classical scaling is an eigenvalue−eigenvector method, but further details will not be discussed here.

Partitioning Fuzzy Cluster Analysis. Cluster analysis permits the study of the dissimilarity matrix as a whole, not only an *m*-dimensional projection of it. Partitioning methods in cluster analysis find the best partition of the *n* objects in *k* clusters. In the method used, the estimation of the objects' membership is carried out by minimization of the objective function[15a]

$$\sum_{v=1}^{k} \frac{\sum_{i,j=1}^{n} u_{iv}^2 u_{jv}^2 \delta_{ij}}{2\sum_{j=1}^{n} u_{jv}^2} \quad (15)$$

where u_{iv} stands for the unknown membership of the object *i* in cluster *v*, and δ_{ij} is the dissimilarity between object *i* and *j*. Fuzzy algorithm details will not be described here, but we note that they provide a more precise description of the objects' membership than the usual hard clustering because they allow for some ambiguity in the data. To measure how hard a fuzzy clustering is, the *normalized Dunn's partition coefficient*[17] is defined as:

$$F_k = \frac{k(\sum_{i=1}^{n}\sum_{v=1}^{k} u_{iv}^2/n) - 1}{k-1} \quad (16)$$

For a completely fuzzy clustering (all $u_{iv} = 1/k$), this coefficient takes on its minimal value 0, whereas the hardest partition (all $u_{iv} = 0$ or 1) yields the maximal value 1. Calculations have been carried out with fuzzy analysis (FANNY) software.[18]

Agglomerative Hierarchical Cluster Analysis. The algorithm constructs a tree-like hierarchy starting with *n* clusters (one for each object), and proceeding by successive fusions until a single cluster is obtained containing all the objects. The method used is based on the *unweighted pair-group average method* of Sokal and Michener,[19] improved by Lance and Williams.[20] The method consists of finding a dissimilarity between two clusters, defined as the average of all the dissimilarities belonging to these clusters:

$$\mathrm{d(R,Q)} = \frac{1}{N_\mathrm{R}N_\mathrm{Q}} \sum_{\substack{i \in \mathrm{R} \\ j \in \mathrm{Q}}} \delta_{ij} \quad (17)$$

where R and Q are the two compared clusters, N_R and N_Q denote their number of objects, and δ_{ij} are the dissimilarities between the objects belonging to clusters R and Q. After each fusion a new dissimilarity matrix is obtained by applying the group average rule (eq 17) to the newly formed clusters. Calculations were carried out with agglomerative nesting (AGNES) software.[18]

5. FLUORO- AND CHLOROSUBSTITUTED METHANES: AN ILLUSTRATIVE EXAMPLE

Fluoro- and chloromethanes are a well-studied set of molecules in the field of quantum similarity.[10,11,21] They have been chosen here as illustrative examples of the comparison between the different MQSM and MQSI. The studied set is made up of the methane and its eight fluoro- and chloro-derivatives: CH_4, CH_3F, CH_3Cl, CH_2F_2, CH_2Cl_2, CHF_3, $CHCl_3$, CF_4, and CCl_4. Density functions and geometries have been carried out using five-function level atomic shell approximation (ASA) fitted densities.[21−23] Geometry was optimized and density functions were normalized to the total number of electrons. Some of the MQSM and MQSI

Table 1. Procrustes Statistic Matrix for the Six MQSM[a]

	O	C	G	TD1	TD8	TD9
O	0	0.21807	$0.85119 \cdot 10^{-1}$	0.51404	0.49543	0.17401
C		0	$0.72595 \cdot 10^{-1}$	0.63551	0.69540	0.42384
G			0	0.56856	0.59041	0.31660
TD1				0	0.28344	0.38225
TD8					0	0.42021
TD9						0

[a] Key: O, overlap MQSM; C, Coulomb MQSM; *G*, gravitational MQSM; TD1, TD8, and TD9, triple-density MQSM with molecule-operators CH_4, CF_4, and CCl_4, respectively.

Table 2. Membership Coefficients for the Six MQSM[a]

MQSM	Cluster 1	Cluster 2	Cluster 3
Overlap	0.5210	0.1139	0.3652
Coulomb	0.8696	0.0500	0.0804
Gravitational	0.9472	0.0181	0.0347
Triple density 1	0.1353	0.6523	0.2124
Triple density 8	0.0573	0.8568	0.0858
Triple density 9	0.0120	0.0109	0.9771

[a] Dissimilarity matrix obtained by Procrustes analysis.

matrixes used in this study have already appeared in a previous paper,[21] and will not be reproduced here.

5.1. Comparison between Molecular Quantum Similarity Measures. As has been described in *Section 2.1.*, there are several definitions of MQSM, depending on which operator is chosen. Four operators were presented (overlap, Coulomb, gravitational, and triple density) yielding four types of MQSM. In the latter case, triple-density MQSM had another degree of freedom; that is, the molecular density function chosen to be the definite positive operator. Here we will consider three molecule-operators: the methane (CH_4) and the completely substituted CF_4 and CCl_4. Thus, we will compare six different MQSM: overlap-like, Coulomb-like, gravitational-like, triple-density with operator CH_4 (that will be denoted by TD1) triple-density with operator CF_4 (TD8), and triple-density with operator CCl_4 (TD9). All these MQSM will be compared by the two methods described in *Section 3*.

Procrustes Statistic. Application of Procrustes analysis to the configurations derived from the different MQSM definitions yields a Procrustes statistic matrix, giving a comparison between them. This matrix is shown in Table 1, and has been analyzed with partitioning fuzzy cluster analysis because classical scaling could only provide two principal coordinates to project the *point-measures*, and essential information can remain hidden in this way.

Table 2 shows the membership coefficients for the six MQSM studied. In a first approach, two-body operators (overlap, Coulomb, and gravitational) can be assigned to cluster 1, TD1 and TD8 can be strongly associated to cluster 2, and finally TD9 itself constitutes the third cluster. Fuzzy clustering, however, allows the extraction of more detailed information on the data structure. Thus, overlap-like MQSM is the most weakly located from all the MQSM, with only 52% membership in cluster 1. It seems like this type of MQSM is located between cluster 1 (52%) and cluster 3 (36.5%), or constitutes itself a fourth cluster. The value of Dunn's coefficient (0.57) confirms this last hypothesis, indicating that Coulomb and gravitational operators are closer to each other than the remaining two-body operator (overlap).

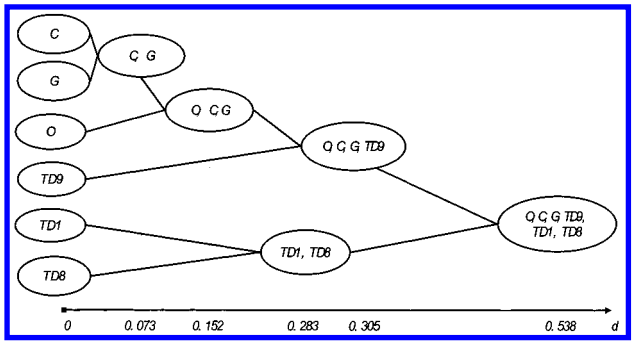

Figure 1. Agglomerative nesting diagram for the six MQSM derived from the Procrustes statistic matrix. Values on the horizontal axis are the dissimilarities between clusters. Key: (*O*) overlap MQSM; (*C*) Coulomb MQSM; (*G*) gravitational MQSM; (*TD1*, *TD8*, and *TD9*) triple-density MQSM with molecule-operators CH_4, CF_4, CCl_4, respectively.

Table 3. Euclidean Distance Matrix for the Six MQSM[a]

	O	C	G	TD1	TD8	TD9
O	0	5.2145	1.9741	3.5967	3.7789	2.3626
C		0	3.5109	5.7813	5.7850	6.5371
G			0	5.0204	4.1716	3.7769
TD1				0	1.3125	3.0285
TD8					0	3.4930
TD9						0

[a] See Table 1 for Key; obtained with the scalar product between standardized matrices.

This proximity does not seem to be in good agreement with the Molecular Quantum Similarity Theory, for reasons described later.

Another remarkable thing is that the three triple-density MQSM cannot be assigned to the same cluster; that is, molecule-operators are structurally so different from each other that they appear as different types of operators. Proximity between triple-density MQSM is mainly related to the structure of the molecule-operator, and these results show how triple density MQSM with large molecule-operators seems to be more similar to two-body operators than the triple-density MQSM with small molecule-operators.

By applying agglomerative hierarchical clustering, some results confirming the conclusions obtained with partitioning fuzzy clustering are found. The obtained linkage chain can be displayed in a dendogram (see Figure 1), which shows the first linkage between Coulomb and gravitational operators, followed by the union of overlap and TD9. On the other hand, TD1 and TD8 are linked together, and finally to the rest of operators.

Proximity Measure between Standarized Matrices. The six MQSM matrixes corresponding to six different operators can be mathematically manipulated as described in *Section 2.2.* to find another similarity measure to compare the existing types of MQSM. The obtained results are shown in Table 3. A principal coordinate analysis is then performed to find a visualization of the point-measures. Taking the first three principal coordinates, the representation obtained is shown in Figure 2.

If we only examine the first PC, all three triple-density MQSM possess similar values, and the rest of the operators are spread rather uniformly. The second PC groups TD9 together with the overlap-like and gravitational-like MQSM, whereas Coulomb-like MQSM remain separated from these

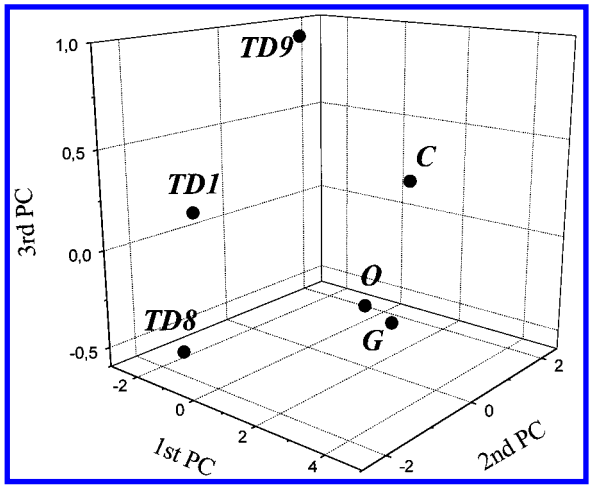

Figure 2. Tridimensional projection of the six-point MQSM using the first three principal coordinates. Key: (*O*) overlap MQSM; (*C*) Coulomb MQSM; (*G*) gravitational MQSM; (*TD1*, *TD8*, and *TD9*) triple-density MQSM with molecule-operators CH_4, CF_4, and CCl_4, respectively.

latter two. Finally, the third PC confirms the rupture between the two-body operators (on one hand, overlap and gravitational, on the other hand, Coulomb), and also confirms the result obtained with Procrustes analysis: TD9 is quite different from the other two triple-density MQSM. Proximity between overlap and gravitational operators is of great interest because they can be considered as even exponents of the same operator: $|\boldsymbol{r}_1 - \boldsymbol{r}_2|^{-n}$, the *coordinate difference operator* (overlap with degree zero and gravitational with degree two). Coulomb, alternatively, represents the least odd exponent of this operator, and, identically, one could define other operators with odd exponents (with degree three, five, etc.). These new operators of odd degrees should be related (closer in a similarity space) between them, and should allow the existence of only two interesting types of coordinate difference operator to be demonstrated; namely, the even and odd exponents of the coordinate difference operator. These two types could be represented by the operators with the least exponent of each class (i.e., overlap and Coulomb operators), although the gravitational operator has been related to the comparison of electrostatic molecular potentials.[24]

These conclusions are in perfect agreement with the hierarchical cluster analysis results summarized in Figure 3. Hierarchical clustering shows the linkage of two-body operators with even exponents (overlap and gravitational), separated from Coulomb. The figure also exhibits the proximity between the small triple-density operators (TD1 and TD8), located far from the larger TD9.

Fuzzy cluster analysis also confirms the aforementioned conclusions. Considering three clusters, Table 4 shows the membership coefficients for the six MQSM. All the results are retrieved, except for TD9, which is included in cluster 1. In any case, a three-cluster division is not good enough ($F_k = 0.52$), and if we performed a four-cluster classification, TD9 would be assigned to the new cluster because of its low membership to cluster 1 (only 54%).

5.2. Comparison between Molecular Quantum Similarity Indices. As has been described in *Section 2.2.*, several possible mathematical manipulations of MQSM exist and produce different MQSI. In particular, five MQSI were

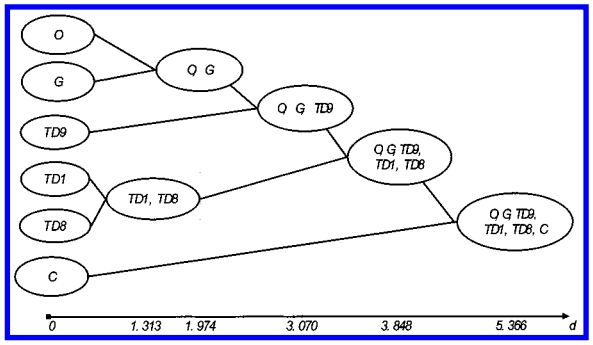

Figure 3. Agglomerative nesting diagram for the six MQSM derived from the scalar product. Values on the horizontal axis are the dissimilarities between clusters. Key: (*O*) overlap MQSM; (*C*) Coulomb MQSM; (*G*) gravitational MQSM; (*TD1*, *TD8*, and *TD9*) triple-density MQSM with molecule-operators CH_4, CF_4, and CCl_4, respectively.

Table 4. Membership Coefficients for the Six MQSM[a]

MQSM	Cluster 1	Cluster 2	Cluster 3
Overlap	0.8721	0.0497	0.0782
Coulomb	0.0041	0.9926	0.0033
Gravitational	0.6087	0.2188	0.1725
Triple density 1	0.0929	0.0527	0.8544
Triple density 8	0.1022	0.0577	0.8400
Triple density 9	0.5394	0.1350	0.3256

[a] Dissimilarity matrix obtained with the scalar product defined in *Section 3.2.*

Table 5. Procrustes Statistic Matrix for the Five MQSI[a]

	D	C	HR	T	P
D	0	0.1950	0.1572	0.1859	0.1637
C		0	$0.1868 \cdot 10^{-1}$	$0.3372 \cdot 10^{-1}$	$0.4529 \cdot 10^{-1}$
HR			0	$0.9684 \cdot 10^{-2}$	$0.1293 \cdot 10^{-1}$
T				0	$0.1503 \cdot 10^{-1}$
P					0

[a] D, Euclidean distance index; C, Carbo index; HR, Hodgkin−Richards index; T, Tanimoto index; and P, Petke index.

defined (Euclidean Distance, Carbó, Hodgkin−Richards, Tanimoto, and Petke). Each one of these MQSI matrixes yields a different configuration of point-molecules, but all of them represent the comparison between the same objects (i.e., the substituted methanes). Therefore, relationships between all these configurations should exist, and Procrustes analysis and the similarity measure defined in *Section 3.2.* will be used to evidence them. To simplify the problem, only overlap MQSM has been used to calculate the MQSI.

Procrustes Statistic. As could be seen in *Section 3.1.*, application of Procrustes analysis to the configurations derived from the different MQSI definitions yields a Procrustes statistic matrix comparing them. This matrix is shown in Table 5. These results were studied again with fuzzy clustering, and Table 6 shows the membership coefficients for the five MQSI.

Although all the MQSI were transformed into D−class indices to be projected with classical scaling, Procrustes analysis clearly retrieved the original division into C−class and D−class indices. Clustering is really hard, as demonstrated by Dunn's coefficient, which takes on a value of 0.87. Unfortunately, cluster analysis does not provide additional information, especially on subdivisions of C−Class MQSI.

Table 6. Membership Coefficents for the Five MQSI[a]

MQSI	Cluster 1	Cluster 2
Euclidean Distance	0.9999	0.0001
Carbó	0.0795	0.9205
Hodgkin−Richards	0.0115	0.9885
Tanimoto	0.0313	0.9687
Petke	0.0568	0.9432

[a] Dissimilarity matrix obtained by Procrustes analysis.

Table 7. Euclidean Distance Matrix for the Five MQSI[a]

	D	C	HR	T	P
D	0	12.9291	13.0830	13.1880	13.2224
C		0	1.0755	1.6384	2.1344
HR			0	1.0712	1.1772
T				0	1.0324
P					0

[a] Obtained with the scalar product between standarized matrices; see Table 5 for Key.

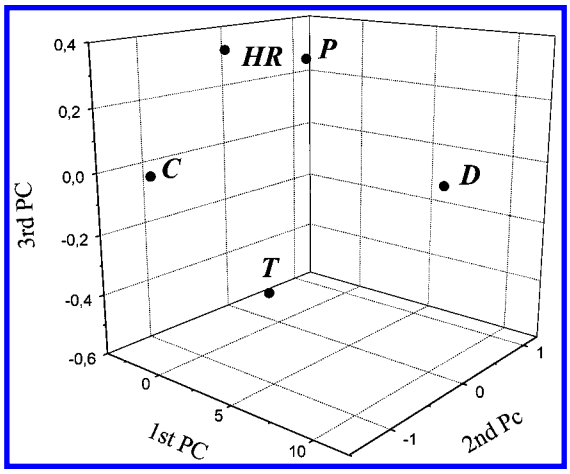

Figure 4. Tridimensional projection of the point-indices using the first three principal coordinates. Key: (*D*) Euclidean distance index; (*C*) Carbó index; (*HR*) Hodgkin−Richards index; (*T*) Tanimoto index; (*P*) Petke index.

Proximity Measure between Standarized Matrices. Table 7 shows the dissimilarity matrix between the five MQSI matrixes obtained with the scalar product defined in *Section 3.2*. A principal coordinate analysis is then performed to project the point-indices in a lower dimensional space. The number of principal coordinates (PCs) chosen is of great importance to have the graphical representation of the point-indices as precise as possible. Taking the first three PCs (Figure 4), the division into D-Class (Euclidean Distance) and C−Class (Carbó, Hodgkin−Richards, Tanimoto, and Petke) indices is clearly visible, despite the fact that MQSI matrixes have been standarized before the comparison. The third PC evidences more subtle relationships between the MQSI. C−Class indices can be divided into two *subclusters*: Carbó, Hodgkin−Richards and Petke indices; and the Tanimoto index. This division can be related to the order of magnitude of these MQSI in relation to the comparison between the same pair of molecules. Thus, the Tanimoto index always gives the lowest C−Class index values[10] of the comparison between any pair of molecules. Further, it is not surprising that Carbó and Hodgkin−Richards indices are closer to each other; in a previous paper,[10] the connection between both of them was demonstrated based on geometrical relationships of the molecular representations generated by the two indices. In Figure 4 the Hodgkin−Richards index seems to be closer to the Petke index than to the Carbó index, in contradiction to this connection. The agglomerative clustering, however, evidences this misinterpretation: the first linkage is between Tanimoto and Petke indices, the next is between Carbó and Hodgkin−Richards, and then the four C−Class indices fusion appears.

Table 8. Membership Coefficients for the Four C−Class Indices[a]

C-Class MQSI	Cluster 1	Cluster 2
Carbó	0.9531	0.0469
Hodgkin−Richards	0.5149	0.4851
Tanimoto	0.2017	0.7983
Petke	0.1336	0.8664

[a] Original dissimilarity matrix obtained with the scalar product defined in *Section 3.2*.

Partitioning fuzzy clustering does not bring forward additional information; it only confirms the hard clustering of the data ($F_k = 0.88$) and the clear division between the two classes of MQSI. Further, if we apply cluster analysis to the C−Class reduced dissimilarity matrix (that is, to the matrix shown in Table 7 without the first column), one obtains the membership coefficients shown in Table 8. This analysis confirms the closeness between the Tanimoto and Petke indices, whereas the Carbó index is assigned to a different cluster and the Hodgkin−Richards index would be located between them. For a more precise appreciation, three clusters would be needed to explain the proximity between the different C−Class MQSI.

6. CONCLUSIONS

The choice of the similarity measures or indices necessary for the study of a molecular set in a determined problem is a difficult task. Many measures and indices have been defined, and some of them are related each other, producing redundant information. To detect these redundancies, a proximity measure between the similarity measures and indices would be of great interest.

In particular, the comparison of the objects belonging to a molecular set using Quantum Similarity Theory has several degrees of freedom, and depending on which definite positive operator or which index transformation is chosen, a different molecular point cloud configuration will be obtained. All these configurations do not produce new information for the knowledge of the similarity between the studied molecular set, and it is reasonable to think that they should be related. Two proximity measures between these different configurations have been described: the Procrustes statistic and a proximity measure involving standarized matrixes-into-vectors. The subsequent data analysis confirmed the conclusions achieved through the direct comparison of the different MQSM and MQSI matrixes, discussed by Carbó et al.[10] Unfortunately, the described methodology and the conclusions derived from it are not general because these similarity measures depend on the quantum system set studied. Thus, we are not comparing MQSM or MQSI, but the application of these MQSM or MQSI in a concrete system; in this case, the substituted methanes.

In reference to MQSM, the comparative study has shown that for triple-density MQSM, the larger the molecule-

operators are, the closer they will be to two-body operators. Further, the proximity between overlap and gravitational operators can be attached to the fact that they can be understood as members of the same operator class; that is, the coordinate difference operator with even exponents.

In reference to MQSI, a clear separation between C−Class and D−Class indices is found. This fact implies that all the possible mathematical manipulations of MQSM can be summarized in two different groups, correlation-like or distance-like transformations, and results obtained using different MQSI belonging to one of these two classes will be almost identical. Going deeply into the study of MQSI, C−Class indices were dealt with separately. Two subclasses were found and are related to the values taken on by the MQSI matrix elements (one subclass includes the Carbó, Hodgkin−Richards and Petke indices, and the other includes the Tanimoto index, which takes on the lowest values). Moreover, the Carbó and Hodgkin−Richards geometrical connection has been retrieved in terms of a considerable closeness in the similarity space.

Finally, it must be noted that this study can be straightforwardly extended to other more general similarity measures and indices provided they produce a configuration in a similarity space 'susceptible' to being compared with the methods discussed here.

ACKNOWLEDGMENT

This work has been partially sponsored by the project SAF 96-0158 of the CICYT. The authors gratefully acknowledge Lluís Amat for providing the fluoro- and chloromethane MQSM matrixes used in this study, and Dr. Emili Besalú for enlightening conversations.

REFERENCES AND NOTES

(1) *Concepts and Applications of Molecular Similarity*; Johnson, M. A., Maggiora, G. M., Eds.; Wiley: New York, 1990.

(2) (a) Carbó, R.; Arnau, J.; Leyda, L. How similar is a molecule to another? An electron density measure of similarity between two molecular structures. *Int. J. Quantum Chem.* **1980**, *17*, 1185−1189. (b) Carbó, R.; Domingo, L. LCAO-MO similarity measures and taxonomy. *Int. J. Quantum Chem.* **1987**, *23*, 517−545 (c) Carbó, R.; Calabuig, B. Molecular quantum similarity measures and *N*-dimensional representation of quantum objects. I. Theoretical foundations. *Int. J. Quantum Chem.* **1992**, *42*, 1681−1693.

(3) Robert, D.; Carbó-Dorca, R. On the extension of quantum similarity to atomic nuclei: Nuclear quantum similarity. *J. Math. Chem*., in press.

(4) Besalú, E.; Carbó, R.; Mestres, J.; Solà, M. Foundations and recent developments in molecular quantum similarity. In *Topics in Current Chemistry*; Sen, K., Ed.; Springer-Verlag: Berlin, 1995.

(5) Carbó, R.; Calabuig, B.; Besalú, E.; Martínez, A. Triple density molecular quantum similarity measures: A general connection between theoretical calculations and experimental results. *Mol. Engineering* **1992**, *2*, 43−64.

(6) (a) Hodgkin, E. E.; Richards, W. G. Molecular similarity based on electrostatic potential and electric field. *Int. J. Quantum Chem.* **1987**, *14*, 105−110. (b) Good, A. C.; Hodgkin, E. E.; Richards, W. G. Utilization of Gaussian functions for rapid evaluation of molecular similarity. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 188−191.

(7) Tou, J. T.; González, R. C. *Pattern Recognition Principles*; Addison-Wesley: Reading, MA, 1974.

(8) Petke, J. D. Cumulative and discrete similarity analysis of electrostatic potentials and fields. *J. Comput. Chem.* **1993**, *14*, 928−933.

(9) Carbó, R.; Besalú, E. Theoretical foundation of quantum molecular similarity. In *Molecular Similarity and Reactivity: From Quantum Chemical to Phenomenological Approaches*; Carbó, R., Ed.; Kluwer Academics: Dordrecht, 1995.

(10) (a) Carbó, R.; Besalú, E.; Amat, Ll.; Fradera, X. On quantum molecular similarity measures (QMSM) and indices (QMSI). *J. Math. Chem.* **1996**, *19*, 47−56. (b) Carbó-Dorca, R.; Besalú, E.; Amat, Ll.; Fradera, X. Quantum molecular similarity measures: Concepts, definitions, and applications to quantitative structure−property relationships. In *Advances in Molecular Similarity*; Carbó-Dorca, R., Mezey, P. G., Eds.; JAI: Greenwich, CT, 1996; Vol. 1, pp 1−42.

(11) Carbó-Dorca, R.; Besalú, E. Extending molecular similarity to energy surfaces: Boltzmann similarity measures and indices. *J. Math. Chem.* **1996**, *20*, 247−261.

(12) Gower, J. C.; Legendre, P. Metric and Euclidean properties of dissimilarity coefficients. *J. Classification* **1986**, *3*, 5−48.

(13) (a) Cox, T. F.; Cox, M. A. A. *Multidimensional Scaling*. Chapman & Hall: London, 1994. (b) Krzanowski, W. J.; Marriott, F. H. C. *Multivariate Analysis*; Edward Arnold: London, 1994; Vol. 1.

(14) (a) de Leeuw, J.; Heiser, W. Theory of multidimensional scaling. In *Handbook of Statistics*; Krishnaiah, P. R., Kanal, L. N., Eds.; North-Holland: Amsterdam, 1982; Vol. 2. (b) Kruskal, J. B.; Wish, M. *Multidimensional Scaling*; Sage: Beverly Hills, 1978. (c) Davison, M. L. *Multidimensional Scaling*; Wiley: New York, 1983.

(15) (a) Kaufman, L.; Rousseeuw, P. J. *Finding Groups in Data: An Introduction to Cluster Analysis*; Wiley: New York, 1990. (b) Krzanowski, W. J.; Marriott, F. H. C. *Multivariate Analysis*; Edward Arnold: London, 1995; Vol. 2. (c) Everitt, B. S. *Cluster Analysis*; Wiley: New York, 1993.

(16) Mardia, K. V.; Kent, J. T.; Bibby, J. M. *Multivariate Analysis*; Academic: London, 1979.

(17) Dunn, J. C. Indices of partition fuzziness and the detection of clusters in large data sets. In *Fuzzy Automata and Decision Processes*; Gupta, M., Ed.; Elsevier: New York, 1976.

(18) Kaufman, L.; Rousseeuw, P. J. CLUSFIND software package. The code includes FANNY and AGNES programs used in this paper, and can be downloaded from the Internet website: http://win-www.uia.ac.be/u/statis/.

(19) Sokal, R. R.; Michener, C. D. A statistical method for evaluating systematic relationships. *University Kansas Sci. Bull.* **1958**, *38*, 1409−1438.

(20) Lance, G. N.; Williams, W. T. A general theory of classificatory sorting strategies: 1. Hierarchical systems. *Comput. J.* **1966**, *9*, 373−380.

(21) Amat, Ll.; Carbó-Dorca, R. Quantum similarity measures under atomic shell approximation: First-order density fitting using elementary Jacobi rotations. *J. Comput. Chem.* **1997**, *18*, 2023−2039.

(22) Constans, P.; Amat, Ll.; Fradera, X.; Carbó-Dorca, R. Quantum molecular similarity measures (QMSM) and the atomic shell approximation (ASA). In *Advances in Molecular Similarity*; Carbó-Dorca, R., Mezey, P. G., Eds.; JAI: Greenwich, CT, 1996; Vol. 1, pp 187−211.

(23) Constans, P.; Carbó, R. Atomic shell approximation: Electron density fitting algorithm restricting coefficients to positive values. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1046−1053.

(24) Carbó, R.; Calabuig, B.; Vera, L.; Besalú, E. Molecular quantum similarity: Theoretical framework, ordering principles, and visualization techniques. *Adv. Quantum Chem.* **1994**, *25*, 253−313.

CI970105U