Quantum Similarity Measures, Molecular Cloud Description, and Structure-Properties Relationships†

Ramon Carbó* and Blanca Calabuig

Institute of Computational Chemistry, University of Girona, Plaça Hospital 6, 17071 Girona, Spain

Received June 12, 1992

The concept of quantum similarity permits us to obtain useful information on the relationships between members of any molecular set and their properties. Quantum molecular similarity proves to be a convenient tool with which to represent molecules as points in an Euclidean space. A projection from an infinite dimensional space to a finite dimensional one, based on principal components of some similarity index matrix taken as molecular coordinates, allows the classification and the visualization of the clustering patterns appearance within a known molecular set. We propose that the basic initial techniques to do this task need to be described within quantum mechanical principles and must obey strict geometrical considerations. After a short discussion on the theoretical structure of quantum similarity measures, there are also described triple density similarity measures, a new way to obtain a projection of the molecular density functions into finite dimensional spaces. Other concepts related to molecular quantum similarity are the description of molecular point clouds and their elements: point-molecules. In order to provide an example, visualization and classification techniques are applied to a molecular family.

1. INTRODUCTION

The accelerated hardware evolution and the development of advanced computational techniques have increasingly influenced the accurate and comprehensive understanding of molecular structure from a theoretical point of view. In this sense, the outburst of molecular similarity theory in the past decade may be viewed as a natural consequence of this scientific endeavor.[1]

Molecular quantum similarity has been defined very recently[2-4] as a new path to conquer a new segment of knowledge, and in order to have a rigorous look over the molecular similarity subject has been observed from a quantum theoretical privileged position. In this way the first branch of a modern discipline, which may be called *quantum molecular engineering*, is subsequently proposed.

Following this idea, in this paper we study the classical definition of molecular quantum similarity in a completely general manner, as well as some new derived concepts. Proceeding along this path, the underlying theory allows us to extend the perspective we already have on this subject.

As a consequence of this new strategy, the developed theory appears to be extremely useful in order to blow up and polish known ideas, and also to analyze new cases in a more comfortable manner, adding a new dimension to the meaning and application of *molecular quantum similarity measures*.

Due to this previous consideration, the present paper consists of four main sections where all the aspects of the theory are shown and discussed. A general description of *quantum similarity* starts the initial part here; the theory in this section is described in order to analyze the application of first principles over any quantum system. A subsection of this first part discusses the differential points which positively split the concept of similarity measure from similarity index. Application of the general formalism to molecular structures constitutes the second part of the discussion. A third section deals with the definition of the novel concepts, based on the previous molecular similarity measures description but attached to *triple density measures*. A fourth chapter is devoted to examining and analyzing some aspects of quantum similarity measures and indices. Finally we present, as an application example, some practical calculations performed by means of two points of view: the classical structure of molecular quantum similarity and triple molecular density measures.

2. QUANTUM SIMILARITY

Reference 5 can be easily chosen as the first published paper where the concept of molecular similarity appears attached to a quantum theoretical viewpoint. According to the spirit of this reference, it is proposed here that quantum similarity measures are to be defined under the outline of quantum mechanical logic and assembled within a correct mathematical framework to preserve as much as possible quantum mechanics principles.[3]

In order to provide a foundation for the previous basic intention, we will suppose that *quantum similarity measures* are completely separated conceptually from *similarity indices*. We will proceed in this way to avoid what seems like a widespread misconception. Similarity measures and indices will be studied here as two completely detached, although related, subjects. This can be firmly based on the fact that similarity indices are nothing more but subsidiary trivial manipulations of similarity measures, which appear in turn to be uniquely defined as the cornerstones of the quantum similarity theoretical building.

If the main purpose of a generally defined similarity measure consists of extracting information from the comparison of two systems, then quantum similarity measures must be based on the main source of quantum information: *density functions*.

pth-order density functions were defined 40 years ago in the well-known Löwdin[6a] and McWeeny[6b] pioneer papers. Density functions, as the ones used in this paper, are obtained upon integration of any electronic quantum system wavefunction square modulus over any particle coordinate except p of them.

The result of this integration process is a function of a $3p$-dimensional variable position vector, $\Gamma(X)$, representing the probability density of finding p system's electrons in an

† A contribution of the Grup de Química Quàntica de l'Institut d'Estudis Catalans.

0095-2338/92/1632-0600$03.00/0

infinitesimal volume element of a p particle position space. The resulting functions, when considered as probability distributions, are ideal candidates to construct similarity measures. Indeed, they are definite nonnegative, continuous, integrable functions, which belong to some infinite-dimensional metric vector space where norms and scalar products between their elements are well defined.

If a known set of quantum systems as objects of study is chosen and every element of the set is essentially distinct of the rest, then the set of density functions for each system, taken in a given state, forms a linearly independent set of functions. Every density function is attached in a one-to-one correspondence to each quantum system, so one can safely suppose that the chosen functional set works as some algebraic representation of the system set. Each density function in turn can be viewed as a point described in an infinite dimensional space and in this manner is manipulated according to the properties of the sets of points in metric spaces.

2.1. Quantum Similarity Measures. Quantum similarity measures constitute the natural starting point of the possible mathematical manipulations which can be made over the set **D** whose elements are density functions. In order to focus the attention on the general structure that envelops the concept of quantum similarity, let us define a coordinate hypervector of p particle positions with the symbol:

$$X = (x_1, x_2, ..., x_p) \tag{1}$$

And let us suppose that the pth-order density functions, as defined by Löwdin,[6a] of two quantum systems I and J, which we represent as $\Gamma_I^{(p)}(X_I)$ and $\Gamma_J^{(p)}(X_2)$, respectively, is known. Assuming also a definite positive operator $\Omega(X_1, X_2)$ of the $2p$ particle coordinate vectors of both systems, then one can define the integral:

$$Z_{IJ}^{(p)}(\Omega) = \int\int \Gamma_I^{(p)}(X_1)\Omega(X_1, X_2)\Gamma_J^{(p)}(X_2)\, dX_1\, dX_2 \tag{2}$$

as the *pth order quantum similarity measure* between two quantum systems I and J.

2.2. Particular Cases. Starting from the general definition given in eq 2, two particular versions of the quantum similarity measure can be immediately defined:

(a) The previous general definition can be easily simplified by using a Dirac's δ function $\delta(X_1 - X_2)$ instead of the definite positive operator $\Omega(X_1, X_2)$. Performing this substitution in eq 2, one arrives at the integral development:

$$\begin{aligned} Z_{IJ}^{(p)}(\delta) &= \int\int \Gamma_I^{(p)}(X_1)\delta(X_1 - X_2)\Gamma_J^{(p)}(X_2)\, dX_1\, dX_2 \\ &= \int \Gamma_I^{(p)}(X)\Gamma_J^{(p)}(X)\, dX \end{aligned} \tag{3}$$

which can be expressed in a convenient simplified notation, dropping all unnecessary indices, like a *density function overlap* by means of the scalar product:

$$Z_{IJ} = \langle \Gamma_I(X) | \Gamma_J(X) \rangle \tag{4}$$

(b) Another possible quantum similarity measure form can be also derived simplifying eq 2 by applying considerations other than these employed in case a, above, but related. Using the one set of particles with a definite positive operator, $\Theta(X_1)\delta(X_1 - X_2)$, instead of the two set of particles, $\Omega(X_1, X_2)$, the following measure is defined, appearing as an alternative of the simpler eq 3:

$$\begin{aligned} Z_{IJ}(\Omega) &= \int\int \Gamma_I(X_1)\Theta(X_1)\delta(X_1 - X_2)|\Gamma_J(X_2)\, dX_1\, dX_2 \\ &= \int \Gamma_I(X)\Theta(X)\Gamma_J(X)\, dX \\ &= \langle \Gamma_I(X) | \Theta(X) | \Gamma_J(X) \rangle \\ &= Z_{IJ}(\Theta) \end{aligned} \tag{5}$$

where, as in eq 4, the indices without a specific function have been omitted for clarity.

These integral measures can be computed by using a couple of system representative functions formed by the same pth-order density function, giving in this case a *quantum self-similarity measure*, which has the following simple integral form according to eq 5 above:

$$Z_{II}(\Theta) = \int \Gamma_I(X)\Theta(X)\Gamma_I(X)\, dX \tag{6}$$

The integral (eq 5) is in fact a general expression of the simplified theory, because the measure defined in eq 3 is nothing more than a particular case of eq 5 when the operator $\Theta(X)$ is substituted by the unit one.

2.3. Similarity Indices. Similarity measures can be readily transformed bearing in mind that they represent some scalar product between two functions belonging to some kind of metric space. These manipulations, as it has been discussed previously, will originate a new kind of auxiliary terms, which will be called *similarity indices*. A clear division must be taken into account then between measures and indices.

From the previously described pth-order quantum similarity measures, a matrix element such as

$$C_{IJ}^{(p)}(\Theta) = Z_{IJ}^{(p)}(\Theta)/[Z_{II}^{(p)}(\Theta)Z_{JJ}^{(p)}(\Theta)]^{-1/2} \tag{7}$$

is defined in the first place as a similarity index.

It will act, remembering a well-known geometrical definition, as the cosine of the angle subtended by the density function pair Γ_I and Γ_J, weighted by the operator Θ. Equation 7 can also be used to compute a normalized form of quantum similarity.

From now on we will suppose implicitly the weighting operator in the similarity measure and index symbols, except in the cases when it seems necessary to stress the nature and presence of the operator Θ, where it will be explicitly written.

Euclidean distances between pth-order density functions can also be written in a straightforward way once the matrix elements defined in eqs 5 and 6 are known. Consider the formula

$$D_{IJ}^{(p)} = [Z_{II}^{(p)} + Z_{JJ}^{(p)} - (Z_{IJ}^{(p)} + Z_{JI}^{(p)})]^{1/2} \tag{8}$$

In this way we have another convenient method of measuring quantitatively the degree of similarity between two quantum systems.

In any case, cosine $C_{IJ}^{(p)}$ and distance $D_{IJ}^{(p)}$ similarity indices are interchangeable. For example, a distance-like index can be defined from $C_{IJ}^{(p)}$, the cosine-like index defined in eq 7:

$$A_{IJ}^{(p)} = \cos^{-1}(C_{IJ}^{(p)}) \tag{9}$$

A reverse technique can be applied to the Euclidean distance index $D_{IJ}^{(p)}$ upon which a cosine-like similarity index can be constructed by means of the simple recipe:

$$M_{IJ}^{(p)} = 1 - D_{IJ}^{(p)}(D_M^{(p)})^{-1} \tag{10}$$

with the aid of the complementary definition

$$D_M^{(p)} = \max(I,J) D_{IJ}^{(p)} \quad (11)$$

Other similarity indices can be derived from the known values of the $\{Z_{IJ}^{(p)}\}$ quantum similarity measures, as, for example, the one described by Hodgkin and Richards[7] in the field of molecular similarity or using the Tanimoto index[8] among many others, but these various alternatives will surely not give in a significant manner, except in very particular cases, more information than the previously discussed standard ones.[3]

3. MOLECULAR QUANTUM SIMILARITY

3.1. General Outline. When dealing with molecules, quantum similarity may obviously be described as molecular. In our laboratory first-order density functions as the density function basis set to compute molecular quantum similarity parameters have been systematically used since the publication of the initial paper on this subject,[5] although higher-order density functions can be employed, as ref 9 shows.

Once it is accepted to use first-order density functions only, attached to each element of the molecular set, one can also invoke the LCAO approximation as a standard procedure to compute them. Let us suppose that for the Ith molecular system an AO basis set, like $\chi = \{\chi_\mu\}$, is selected and that the system's first-order density matrix $\mathbf{D}_I = \{D_{I,\mu\nu}\}$ is also known.

First-order density functions in the LCAO-MO framework can easily be written as[6]

$$D(X) = \sum_\mu \sum_\nu D_{\mu\nu} \phi_\mu \phi_\nu \quad (12)$$

A *quantum similarity* measure constructed by means of eq 2, when first-order density functions as defined in eq 12 are involved, can be generally built as

$$Z_{IJ}(\Theta) = \int\int D_I(X_1)\Theta(X_1, X_2) D_J(X_2)\, dX_1\, dX_2 \quad (13)$$

where $\Theta(X_1, X_2)$, is an Hermitean operator depending on two-particle coordinates. The most immediate choice for this two-particle operator is, once more, the already mentioned Dirac δ function, $\delta(X_1 - X_2)$. Within this choice, eq 13 transforms obviously into the integral shown in eq 4.

Also Coulomb operators $\Theta(X_1, X_2) = |X_1 - X_2|^{-1}$ can easily be used in eq 13 to provide, along with eq 4, the most usual kinds of quantum similarity measures: Overlap-like and Coulomb-like, respectively.

3.2. Quantum Theoretical Interpretation of Molecular Quantum Similarity. (a) When observing eq 4 and, in general, the definition provided by eq 13 with the Θ operator chosen as the unit one, we can study the quantum similarity $\{Z_{IJ}\}$ integrals from the quantum theoretical point of view.

Then, pth-order density functions $\Gamma(p)(X)$, can also be considered as Hermitean definite positive p-electron operators. In this manner, according to quantum mechanical principles,[10] they are considered operator representations of some kind of the observable system, depending only on the positions of the system's p particles.

In particular, when dealing with arbitrary system density functions $\Gamma(X)$, we can write the following equality sequence, when eq 4 is analyzed:

$$Z_{IJ} = \langle J|\Gamma_I(X)|J\rangle = \langle \Gamma_I(X)\rangle_J = \langle I|\Gamma_J(X)|I\rangle = \langle \Gamma_J(X)\rangle_I \quad (14)$$

where $|I\rangle$ and $|J\rangle$ are the wavefunctions from which the density functions Γ_I and Γ_J are constructed.

The sequence of equalities (eq 14) above suggests that the integral Z_{IJ} can be interpreted as the *expectation value* of the *density operator* $\Gamma_I(x)$ with respect to the *density function* $\Gamma_J(x)$. The same can be said when exchanging the indices I and J, that is, the roles of both density functions.

When two different quantum states or systems are involved, the proposed natural interpretation, invoking quantum mechanical principles again, provides *identical* expectation values to *both* densities considered as operators or functions in turn. On the contrary, quantum self-similarity has then a unique possible interpretation as the expectation value of the density operator, computed within the same density function.

(b) When positive definite weighting operators other than unity are present in the body of quantum similarity measures, the quantum theoretical interpretation can be built up in some cases on the same grounds as it has been above. For instance, when a Coulomb operator is used, one can say that the expectation value of the electrostatic potential of a molecular source regarding the density function of another molecule remains invariant when changing molecular roles, that is, density functions. This is due to the already commented on quantum chemical theoretical background which *attaches* to each molecular state a *unique* density function.

3.3. Approximations and Computation of Similarity Integrals. (a) In molecular quantum similarity computations, it seems also compulsory to study the case of first-order density functions, as it is easy to obtain the information from the output of the usual available programs. Some problems, whose nature was already explained in some previous papers on this subject,[4,5,13b] arise when computation of the similarity integrals over AO is needed. This computational trouble depends on the nature of AOs, the definite positive operator employed into the quantum similarity measure, and the involved integrals which are equivalent in number to the evaluation of an interaction energy between both compared systems. Recently, various authors have proposed other computational approaches to this problem.[14]

(b) A convenient elementary algorithm within a first-order density function choice can be easily deduced, as discussed in refs 5 and 13c. The expression given in eq 13 can be transformed into a CNDO-like one, invoking ZDO overlap first, followed by an average over AOs on the same atom. After these manipulations we obtain

$$Z_{IJ}(\Theta) \approx \sum_{A\in I} \sum_{B\in J} Q_{A,I} Q_{B,J} \int\int |S_A(X_1)|^2 \Theta(X_1, X_2) |S_B(X_2)|^2 \times dX_1\, dX_2 \quad (15)$$

where $\{Q_{A,I}, Q_{B,J}\}$ are Mulliken gross atomic populations on atom A of system I and B of system J, respectively, or any alternative valid form defined as an atomic charge. Reference 12 can be used in order to find a recent definition of charge which may be employed in a safe way on this and any other context of quantum similarity. $\{S_A, S_B\}$ are s-type AOs centered on centers A or B and taken to be the outermost shell s orbital of each atom.

(c) Equation 15 may suggest a possible general form of a simplified molecular quantum similarity measure. Collecting gross atomic populations and ns orbital densities, forming some sort of atomic density, as follows:

$$R_{A,I}(X) = Q_{A,I} |S_A(X)|^2 \quad (16)$$

one can write eq 15 like

$$Z_{IJ}(\Theta) \approx \sum_{A\epsilon I}\sum_{B\epsilon J}\int\int R_{A,I}(X_1)\Theta(X_1, X_2)R_{B,J}(X_2)\,dX_1\,dX_2 \tag{17}$$

It can be seen how in this form the atomic density function sets $\{R_{A,I}\}$ and $\{R_{B,J}\}$ can be easily expressed by means of other functional forms, different from the ones proposed in eq 16. For example, the exact expression (eq 13) will be obtained when using atomic density functions in eq 17 constructed from the partition of first-order density (eq 12):

$$R_{A,I} = \sum_{\mu\epsilon A}\sum_{K\epsilon L}\sum_{\nu\epsilon K} D^{AK}_{\mu\nu;I}\chi^{A}_{\mu}\chi^{K}_{\nu} \tag{18}$$

and some equivalent definition for $R_{R,J}$.

(d) Other possible simplifications of quantum similarity measures, closer to the exact result than approximation 16, are described in ref 3.

4. TRIPLE DENSITY MOLECULAR QUANTUM SIMILARITY MEASURES

4.1. Definition of Triple Density Measures. The quantum interpretation given in the discussion around the meaning of eq 14 with respect to similarity measures, as defined in eq 5, may also suggest the use of the same function-operator duality which bears density functions, allowing in this manner the *p*th-order density functions to be employed as weighting operators. A new rule can be envisaged defining a *triple density quantum similarity measure*, substituting in eq 5 the positive definite operator $\Theta(X)$ by another appropriate *p*th-order density function $\Gamma_C(X)$. That is, one can construct, in general, a new similarity measure integral having the form

$$Z_{AB,C} = \int \Gamma_A(X)\Gamma_C(X)\Gamma_B(X)\,dX = \langle \Gamma_A(X)|\Gamma_C(X)|\Gamma_B(X)\rangle \tag{19}$$

The simplified CNDO scheme to compute quantum similarity previously described in section 3.5(b) above, according to eqs 16 and 17, appears as a very convenient computational structure when triple density measures are to be evaluated. In this case, eq 19, if function 16 is used to represent atomic densities, will have the following structure:

$$Z_{PQ,I} \approx \sum_{A\epsilon P}\sum_{C\epsilon I}\sum_{B\epsilon Q} Q_{A;P}Q_{C;I}Q_{B;Q}\int |S_A(r)|^2|S_C(r)|^2|S_B(r)|^2\,dr \tag{20}$$

It is easy to see which kind of integrals do appear when the first-order density function, as defined in eq 12, is used in the triple density similarity measure as written in eq 19. There emerges some kind of overlap integral involving six AO centered, in the worse situation, on six different atomic sites; for more details see ref 4. This kind of integral is a trivial extension of those appearing in the usual quantum similarity measures, already described. The needed basic integrals may be immediately computed if GTO functions are used, but the computational scheme becomes less simple when exponential type functions, ETO, are involved. Some discussion on the triple density measure basic integral forms are also analyzed in ref 4. In any case, however, these integrals will represent a heavy ab initio burden. The associated computational effort can be alleviated by realizing that quantum similarity integrals are adequate for parallel and vector programming algorithms.

The interpretative explanation of the triple density measure can be made in the same terms as in section 3.2 above by means of the application of quantum mechanical principles. In fact, such a measure, as given in eq 19, corresponds to the matrix element representation of the density function Γ_C, considered as an operator, in some basis set made of the density functions where the couple $\{\Gamma_A, \Gamma_B\}$ belongs. The consequences of this point of view are discussed next.

4.2. Practical Interpretation of Triple Density Similarity Measures over a Set of Density Functions. 4.2.1. Molecular and Density Sets. From a practical point of view, the use of triple density measures must be considered in a context where the knowledge of a set of molecules, **M**, should be well-established. The molecular density functions forming the set **D**, all of the same arbitrarily chosen order, are supposed to be also available. Both sets are constructed in such a way that there is an implicit one-to-one correspondence between their elements.

Each element of the set **D** can be represented as a matrix using in turn all the elements of **D** as a basis set, through triple density measures defined in eq 19. This means that one can define a symmetric matrix $\mathbf{Z}_I = \{Z_{PQ,I}\}$ connected to each element belonging to the density set **D**. This is also the same to say that to each molecule *I* of the set **M** there can be attached a matrix $\mathbf{Z}_I$, which can be uniquely defined.

4.2.2. Algebraic Interpretation. From the algebraic point of view, the density representation matrices, $\mathbf{Z}_I$, can also be viewed as the density function coordinates defined in a vector space Θ_N of finite dimension $N = (n^2 + n)/2$, where $n = \#(\mathbf{M})$ is the number of molecules in the set **M**. Such a finite dimensional space may be considered forming the isomorphic representation of each density function throughout the basis set of the available pairs of density functions belonging to the set **D**. Thus, each density function, represented in such a way, may be seen as a projection from the infinite dimensional space where it belongs into a finite dimensional Euclidean space.

By construction, each molecule in the set **M** has a unique attached density function in the set **D** of the density functions and vice versa. This is equivalent to admit that the matrix $\mathbf{Z}_I$, constructed using triple density measures with the *I*th density function as weighting operator, can be considered as the matrix representation of the associated *I*th molecule in the set **M**. Then, as a result of this mathematical process, a molecule can be represented in some finite dimensional vector space as a point: a *point-molecule*.

The set of point-molecules $\Theta_N = \{Z_I; (I = I, n)\}$ can be submitted to further mathematical manipulations in order to quickly compute proximity information patterns or any other desired kind of relationship between the elements of Θ_N. This element can be transferred to the ones of the set **M** in the same way that was discussed for molecular similarity measures in general. In this sense, we call *molecular point clouds* this kind of Euclidean vector space representations of molecular sets.

Another interesting particularity to be noted, with regard to the algebraic structure attached to triple density measures, is the influence on the molecular point cloud Θ_N that the increase of information knowledge over the sets **M** and **D** has. The accretion of information on the set **M** can be produced by simply adding to it at least one new molecular structure. In this way, the cardinality of **D** is increased accordingly by one, but the dimension of the molecular point cloud space Θ_N is increased in increments of $n + 1$. This happens when redundant matrix elements are ignored. All the vectors, the point-molecules, are now represented by symmetric matrices where a new row and column must be added and one or the other possess redundant elements. In relation to this, the

appearance of a new matrix in Θ_N must be considered attached to the added structure, representing the new weighting density function, incrementing in this way by one the set of available point-molecules. Consequently to the vector space unit, dimension variation corresponds, as a response, to some expansion of the molecular point cloud degrees of freedom.

4.2.3. Reproducibility. A final detail must be given concerning the relative atomic coordinate axis orientation of each molecule in the set **M** before performing the evaluation of the triple density similarity integrals. The same problem has been already discussed in previous quantum molecular similarity calculations.[5,13] In this paper we have chosen the following procedure, which has been systematically applied to every element of the molecular set. Each molecule from the set **M** is represented with its coordinates referred to the center of charges. Later on, quadrupole principal components of each structure are computed, and the translated coordinates are transformed in such a manner as to match quadrupole principal directions with molecular axis. Triple density measures are computed upon this homogeneous molecular coordinate set.

This algorithm is sufficient to make reproducible the results of any calculation of this sort and eliminates the ambiguity in the molecular axes definition of one molecule with respect to the other two present in the integral measure.

5. SOME RELEVANT ASPECTS RELATED TO SIMILARITY MEASURES

Besides everything that has been said so far, it is interesting to make clear some points related to similarity measures which have not been published yet.

(a) First, it is interesting to study a conjecture which can be made extensible to the general techniques of quantum similarity measures. Given an n-object set **M**, defined in an infinite dimensional space, V_∞, there must exist a finite dimensional projection $P_d(\mathbf{M}) = \mathbf{S}$ of the set **M**, over the subspace $V_d \subset V_\infty$, such that an *optimal visualization* of the set **M** could be carried out through a later rotation projection: $R_v(S) = \mathbf{U} \subset V_v$, performed over a two- or three-dimensional *visual space*.

(b) The term *optimal visualization* may have several meanings, and it is not uniquely defined since the point of view from which one can observe a projected molecular point cloud can be considered optimal according to several criteria, such as (1) Distances between the points are maximized. (2) The order of the objects fits in the best way to the order of some molecular property. (3) Some subset of the objects is revealed according to certain similarity conditions. (4) The form of a tree or a graph constructed over the elements of **M** appears completely developed.

(c) It must be taken into account that the information on similarity measures between the elements of the set **M** *always* can be transformed into a *similarity matrix* formed by elements constructed through some similarity index. From such a matrix it can always be constructed an $(n \times n)$ *coordinate matrix* associated with the elements of **M**; for example, computing the similarity matrix *principal components*. These coordinates, which always can be built up to represent any element of the set **M**, could be called *homogeneous, standard, characteristic, or reduced* since any molecular set can be potentially transformed in such a way and so be compared to any other system that has been mathematically manipulated in such a manner.

(d) Then, it is easy to see that we have arrived at a broad description of the molecular quantum similarity problem. Any set of objects *can be made homogeneous* through the computation of

(1) a similarity measure performed in the original space in which the objects are defined.

(2) a similarity matrix through the use of a similarity index transformation of the initial similarity measure. The matrix must be positive definite.

(3) the eigenvectors, principal components, of the similarity matrix.

The eigenvectors obtained by diagonalizing the similarity matrix can act as *homogeneous coordinates* of the initial objects.

(e) From the previous discussion, it may be deduced that any set of object coordinates can be homogenized. With this term we understand that the *global description* of any set of objects can be standardized according to paragraph d. This transports the problem toward the definition of what must be understood by a *description* of a set of objects.

A set of objects **M** may be considered *described* through a set of vectors **U**, when there are such a vector space V and a subset $\mathbf{U} \subseteq \mathbf{V}$ that a one-to-one correspondence between **M** and **U** exists.

Table I. Frog Muscle Activity Inhibition

compd	log (MBC)	compd	log (MBC)
trichloromethane	1.500	ethoxyethane	1.930
methanol	3.090	toluene	1.000
ethanol	3.750	(hydroxymethyl)benzene	1.300
1-propanol	2.400	phenol	1.000
2-propanol	2.550	2-isopropyl-5-methylphenol	−0.520
1-butanol	1.780	aniline	1.300
1-pentanol	1.200	nitrobenzene	0.470
1-hexanol	0.560	pirydine	1.770
1-heptanol	0.200	2-hydroxynaphthalene	0.000
1-octanol	−0.016	quinoline	0.300
propanone	2.600		

6. PRACTICAL RESULTS

Once the numerical representation of a molecular set on some vectorial space has been obtained, that is, when the molecular set can be considered described, one can apply a large variety of numerical analysis and visualization techniques to get a multifaceted image of any molecular point cloud.

In this section the similarity analysis of a molecular family as a particular example is presented to make a comparison between the two possible methodologies of molecular similarity measures previously described.

All computations have been made using MOLSIMIL-90 program, a new version of a published one conceived to perform molecular quantum similarity,[13c] and TRIDENT,[4] a program where triple density measure ideas are implemented.

Visual representations are made by means of HYPERCLOUD, a set of graphic design facilities developed in our laboratory.[3,13] HYPERCLOUD visual results can be presented in a variety of formats: printers, color screens. They may be captured and stored in computer files for further use or directly shot into film slides or photographs, even in video tapes.

A molecular set of organic substances which have frog muscle activity inhibition effects[15] has been studied here. Names and activities of these molecules, expressed as log (MBC) and given in millimoles per liter, are given in Table I.

Biological activity values of the studied molecules have been distributed in two classes, associating a different graphic symbol to each one when drawn, according to its activity value.

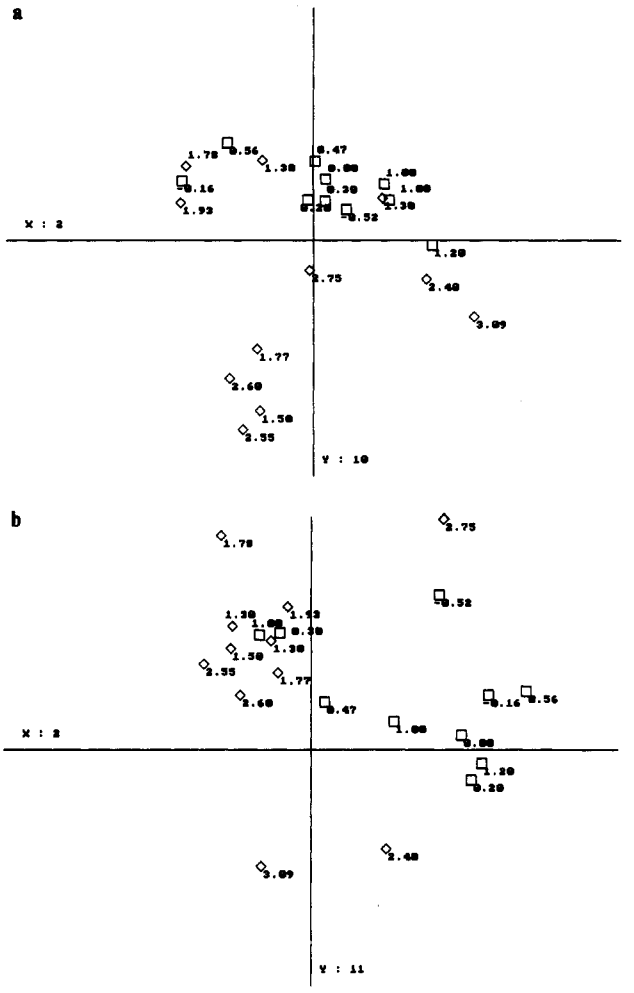

Figure 1. Visualization of a molecular point cloud. The original 28-dimensional space is transformed into an optimal plane by means of principal components manipulation. (a) Similarity measure data. (b) Triple density similarity data.

Figure 1a represents the projection over an optimum bidimensional plane[13d] made of the principal components 2 and 10 of the molecular point cloud similarity matrix, obtained after the projection of the 28-dimensional space over a four-dimensional subspace. Principal components are constructed in turn with the similarity index eigenvectors calculated from MOLSIMIL-90 similarity measures. Figure 1b presents the molecular point cloud projection to a plane defined by the principal components 2 and 11, obtained in the same way as before but from the TRIDENT procedure results. In both panels we can observe the fact that molecules whose activity belongs to the same activity class gather in the same zone of each graphical representation.

In Figure 2, panels a and b, Kruskal trees have been automatically constructed through another HYPERCLOUD procedure, using the results obtained with MOLSIMIL-90 and TRIDENT, respectively, for the previously described molecular set. The Kruskal tree[5] is a graph in which each molecule is represented linked by a line to the most similar molecules. Lines are depicted depending on a Euclidean distance similarity index in this case, as described in eq 8. In these figures, molecules of the studied set are situated on the vertex of a five-dimensional hypercube, which for the sake of clarity has not been drawn. In general, the hypercube dimension must be estimated in such a way that the resulting vertex number becomes equal to or higher than the number of molecules to represent.[2-4]

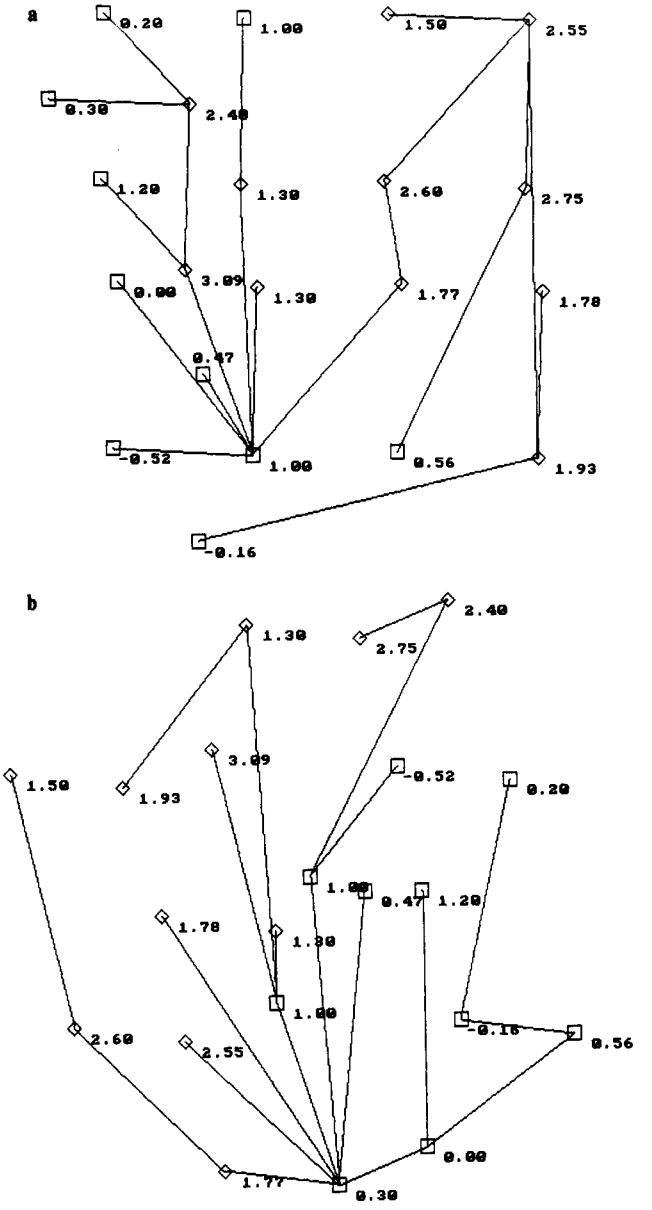

Figure 2. Kruskal tree representations of a molecular point cloud, drawn by putting the point-molecules in the vertices of a hypercube. (a) Similarity measure data. (b) Triple density similarity data.

7. CONCLUSIONS

The elements for the computation of similarity measures from a quantum theoretical point of view have been defined, splitting this concept from similarity index definitions.

Quantum similarity measures and the concepts related to them have been used in a practical form for the development of MOLSIMIL-90 and TRIDENT programs. The visualization of the results obtained, through a graphic programs package, HYPERCLOUD, generates a supplementary interest added to the similarity measures. The search of subsets within the cloud of point-molecules, employing these techniques, becomes an excellent tool in order to study from another point of view the old problem related to structure–activity and structure–properties relationships.

Taking into account the solid theoretical development and the presentation of the results obtained through MOLSIMIL-90 and TRIDENT, we can say that both kinds of computations are complementary, giving each one a particular perception of the similarity relationships between the elements of the analyzed set.

8. COROLLARY

All the material published until now referring to quantum molecular similarity, summarized in this paper, opens in our opinion a modern and suggestive perspective to the endless tasks associated to the classification and visualization of chemical species considered as a whole.

Using a current terminology, the point-molecules and their collections, the molecular point clouds, are not only currently generated by means of a computer but also can be manipulated through a machine until its perception from the human side is reached, materializing somehow and somewhere in the so called *cyberspace*.[16]

At the present time we want to generalize the actual concept of *virtual reality*,[17] which can be sketchy described as a tridimensional environment, generated within the cyberspace, which can be observed by the on line potential users *as something as real as possible.*

The schemes of computation, projection, and visualization of any molecular point clouds, contained in *n*-dimensional cybernetic spaces, can be considered as an initial stage for the creation, manipulation, and observation in *real time* of what is here tentatively named using the term *virtual hyperreality.*

ACKNOWLEDGMENT

This work has been supported by the Spanish Ministerio de Industria y Energía under the Programa Nacional de Investigación y Desarrollo Farmacéuticos through Grant FAR 88-0617 and by the Generalitat de Catalunya through QFN 91-4206 under the Química Fina programme.

REFERENCES AND NOTES

(1) Johnson, M. A., Maggiora, G., Eds.; *Concepts and Applications of Molecular Similarity*; John Wiley & Sons: New York, 1990.
(2) Carbó, R.; Calabuig, B. Quantum Similarity. In S. Fraga (ed.) *Structure, Interactions and Reactivity*; Fraga, S., Ed.; Elsevier: Amsterdam, 1992.
(3) (a) 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. (b) Carbó, R.; Calabuig, B. Molecular Quantum Similarity Measures and *N*-Dimensional Representation of Quantum Objects. II. Practical Applications. *Int. J. Quantum Chem.* **1992**, *42*, 1695–1709.
(4) Carbó, R.; Calabuig, B.; Besalú, E.; Martínez, A. Triple Density Molecular Quantum Similarity Measures: A General Connection between Theoretical Calculations and Experimental Results. Submitted for publication in *Mol. Eng.*
(5) Carbó, R.; Arnau, M.; 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.
(6) (a) Löwdin, P. O. Quantum Theory of Many-Particle Systems. I. Physical Interpretations by Means of Density Matrices, Natural Spin-Orbitals, and Convergence Problems in the Method of Configurational Interaction. *Phys. Rev.* **1955**, *97*, 1474–1489. (b) McWeeny, R. The Density Matrix in Many-Electron Quantum Mechanics. I. Generalized Product Functions. Factorization and Physical Interpretation of the Density Matrices. *Proc. R. Soc. London, A* **1959**, *253*, 242–259.
(7) Hodgkin, E. E.; Richards, W. G. Molecular Similarity Based on Electrostatic Potential and Electric Field. *Int. J. Quantum Chem.* **1987**, *14*, 105–110.
(8) Carbó, R.; Domingo, Ll. *Algebra Matricial y Lineal*; Serie Schaum; McGraw-Hill: Madrid, 1987.
(9) Ponec, R.; Strnad, M. Electron Correlation in Pericyclic Reactivity: A Similarity Approach. *Int. J. Quantum Chem.* **1992**, *42*, 501–508.
(10) (a) Eyring, H.; Walter, J.; Kimball, G. E. *Quantum Chemistry*; John Wiley & Sons: New York, 1948. (b) von Neumann, J. *Mathematical Foundations of Quantum Mechanics*; Princeton University Press: Princeton, 1955. (c) Pauling, L.; Wilson, E. Bright, Jr. *Introduction to Quantum Mechanics with Applications to Chemistry*; Dover Publications: New York, 1985.
(11) Mulliken, R. S. Electronic Population Analysis on LCAO-MO Molecular Wave Functions. *J. Chem. Phys.* **1955**, *23*, 1833–1840, 1841–1846, 2338–2342, 2343–2346.
(12) Huzinaga, S.; Sakai, Y.; Miyoshi, E.; Narita, E. Extended Mulliken Electron Population Analysis. *J. Chem. Phys.* **1990**, *93*, 3319–3325.
(13) (a) Carbó, R.; Arnau, C. Molecular Engineering: A General Approach to QSAR. In *Medicinal Chemistry Advances*; de las Heras, F. G., Vega, S., Eds.; Pergamon Press: Oxford, 1981; pp 85–96. (b) Carbó, R.; Domingo, Ll. LCAO-MO Similarity Measures and Taxonomy. *Int. J. Quantum Chem.* **1987**, *23*, 517–543. (c) Carbó, R.; Calabuig, B. MOLSIMIL-88: Molecular Similarity Calculations Using a CNDO-like Approximation. *Comp. Phys. Commun.* **1989**, *55*, 117–126. (d) Carbó, R.; Calabuig, B. Molecular Similarity and Quantum Chemistry. Reference 1, Chapter 6, pp 147–171. (e) Carbó, R.; Calabuig, B. Quantum Molecular Similarity Measures and the *N*-Dimensional Representation of a Molecular Set: Phenyldimethylthiazines. *J. Mol. Struct. (Theochem)* **1992**, *254*, 517–531.
(14) (a) Cooper, D. L.; Allan, N. L. Molecular Dissimilarity: A Momentum-Space Criterion. *J. Am. Chem. Soc.* **1992**, *114*, 4773–4776. (b) Cioslowski, J.; Fleischmann, E. D. Assessing Molecular Similarity from Results of ab Initio Electronic Structure Calculations. *J. Am. Chem. Soc.* **1991**, *113*, 64–67.
(15) Agin, D.; Hersh, L.; Holtzmann, D. *Proc. Natl. Acad. Sci. U.S.A.* **1965**, *53*, 952.
(16) See, for example: *Sci. Am.* **1991**, *Sept* (special issue).
(17) See, for example: Davis, D. B. Visualization's New Breed. *Comput. Graphics World* **1991**, *June*, 44–52.