J Math Chem (2013) 51:1569–1583
DOI 10.1007/s10910-013-0165-4

ORIGINAL PAPER

Particle coordinates and discrete molecular description: a geometric point of view on a twofold dimensionality environment

Ramon Carbó-Dorca

Received: 17 December 2012 / Accepted: 14 March 2013 / Published online: 21 March 2013

Abstract In the present study a specific kind of widespread mathematical objects: descriptor matrices are defined and studied. These are matrices connected with several problems concerning many fields of interest in theoretical chemistry, classical and quantum mechanics, or quantitative structure-properties relations. The twofold dimensionality structure of descriptor matrices is analyzed and the properties of descriptor matrices are also disclosed with respect origin shifts and rotations. A tensor to study schematically the three dimensional nature of many particle structures, the characteristic form tensor is defined. The construction of similarity matrices from descriptor matrices and the connection with quantum similarity are finally discussed.

Keywords Twofold dimensionality · Many particle position spaces · Conformational molecular spaces · QSPR descriptor spaces · Origin shift of vectors sets · Complexes · Chemical object spaces · Descriptor matrices · Dimensionality paradox · Characteristic form tensor · Similarity matrices · Quantum similarity · Quantum similarity matrices

1 Introduction

Several studies concerning the so-called dimensionality paradox in the field of quantitative structure-properties relations (QSPR) and the theory and practice of quantum QSPR (QQSPR) have been recently published [1,2]. At the same time, some printed theoretical papers also deal about the origin shift technique applied within quantum object sets (QOS) [3–6], with an obvious extension involving classical descriptor parameters, see for example reference [2]. In the two mentioned work series, the theoretical structure of quantum similarity (QS) in general and molecular QS (MQS) in particular

R. Carbó-Dorca (✉)
Institut de Química Computacional, Universitat de Girona, 17071 Gerona, Catalonia, Spain
e-mail: quantumqsar@hotmail.com

[4,7], have a relevant role. Even more recently, a discussion over the definition of chemical and molecular spaces also appeared [8], followed by an attempt to describe the concept of molecular universes [9] as a way to clarify as much as possible the basic background of the QSPR subject.

The present work is related to all the quoted direction lines and will try to advance in the analysis of the description based on a geometric viewpoint of both position of particles and discrete representation of molecular sets. This project will be performed by means of some general kind the matrices which will be called *descriptor* matrices. In order to achieve this goal, the present contribution will be organized as follows. The statement of the mathematical background will be given first as a way to provide basic definitions and mathematical notation. Next the geometrical point of view of the problem will be discussed; in this part the origin shift technique and the so-called dimensionality paradox will be introduced. In the next section origin shifts will be exhaustively applied into the object spaces by using vertex shifts and the relevant results discussed. In the next section origin shifts using convex combinations of vertex coordinates will be discussed; origin shift via the centroid of the vertex set will be also studied as a particular case. In the next section will be discussed the many dimensional rotations of the involved descriptor matrix elements. A following section will discuss the possibility to construct square matrices from the descriptor matrices; based in the twofold dimensionality the adequate candidates are the corresponding Gram matrices associated to the vectors of both associated spaces. Next section will be finally devoted to the construction of similarity matrices by means of descriptor matrices and its connection with quantum similarity matrices.

2 Statement of the mathematical background

This first section describes the data structure which will be met in the problems discussed hereafter. It is a trivial matter to set up such data formalism, but it is initially presented here as a way to also set up the notation which will be used through the present study.

The ordering of the involved objects, which will be described in several ways ending at the same formal structure, is arbitrary and it is supposed that do not have any influence in the properties of the matrices which will be discussed henceforth.

Two apparently diverse situations can be present.[1] One can consider first three-dimensional position space location of N equal or different particle sets, as appear in Monte Carlo procedures or alternatively when setting molecular conformations respectively. In second term one can be using some kind of M-dimensional spaces as the container of N molecular structures, when constructing the discrete representation of molecular sets by means of the so-called descriptor parameters.

Then, in general it can be said that these aforementioned situations as a whole correspond to the construction of formally similar mathematical objects. This is so as

[1] In fact, there must be also considered present an alternative parallel description in dual space, involving the transposes of all vectors and matrices which are employed in this work. The twofold dimensions appearing in the main discussion will be therefore reversed looking at this dual case. However, such issue will not be discussed here, in order to keep the formalism as simple as possible.

these problems end up with the construction of some $(M \times N)$-dimensional *descriptor* matrix $\mathbf{X}$.

The setup characteristics of the involved description space dimensions in the cases of particle position, precludes the use of the left dimension parameter: $M = 3$, because three-dimensional space is involved; while at the same time usually the right dimension parameter becomes: $N \gg 3$, with the exception within the conformational studies of small molecules, as di- or triatomic structures. However, in the last case of molecular discretization problems, dimensions at the left (row) and right (column) sides habitually are related by the distinctive property: $3 \gg M \gg N$.

In any case one can speak about a *twofold* dimension or dimensionality, characterized by the respective choices of M and N, according to the nature of the represented problem.

The meaning of the elements inside the $(M \times N)$ descriptor matrices is quite obvious in all the mentioned cases. Noting provisionally an appropriate descriptor matrix and its elements as:

$$\mathbf{X} = \{X_{IJ} \,|\, I = 1, M \;;\; J = 1, N\} \tag{1}$$

such a matrix can be decomposed in a set of N *column*[2] vectors:

$$C = \{|\mathbf{c}_J\rangle \,|\, J = 1, N\} \subset \mathrm{V}_M \wedge \forall J : |\mathbf{c}_J\rangle = \{X_{IJ} \,|\, I = 1, M\} \tag{2}$$

which is a vector subset belonging to some M-dimensional column vector space. Let me call these vectors *object* vectors and the space where they belong the *object* space or space of the objects.

However, the same descriptor matrix can be also supposed alternatively partitioned; now using M *row* vectors as:

$$\mathbf{X} = \begin{pmatrix} \langle \mathbf{f}_1| \\ \langle \mathbf{f}_2| \\ \vdots \\ \langle \mathbf{f}_M| \end{pmatrix} \rightarrow F = \{\langle \mathbf{f}_I| \,|\, I = 1, M\} \subset \mathrm{V}_N \wedge \forall I : \langle \mathbf{f}_I| = \{X_{IJ} \,|\, J = 1, N\} \tag{3}$$

and in such a way that the resultant vector set belongs to an N-dimensional row vector space. Let me call these vectors *descriptor* vectors and the space where they belong the *descriptor* space, description space or space of the descriptors.

Therefore, insisting in what it has been earlier commented, a common attribute to all possible description scenarios as these here analyzed, is the existence in every problem of an attached twofold dimension column–row vector sets, belonging respectively to object and descriptor spaces.

[2] Here the Dirac's notation is used for column vectors, which will be written using a ket symbol: $|\mathbf{a}\rangle$, while row vectors will be noted by a bra symbol $\langle \mathbf{a}|$. Both symbols correspond to the transpose one from the other. As the field of reference along this study will be the real field, then there is no conjugation involved.

With these preliminary definitions and comments one can proceed to study the general properties associated to the descriptor matrices.

3 The geometrical point of view

Observing the descriptor matrix definition in Eq. (1) and its partitions (2) and (3), it clearly appears that the twofold dimension partitions of this kind describe two types of sets made of polyhedral vertices in the corresponding attached spaces, differing slightly according the described situations.

3.1 Position problems

In all the cases of three-dimensional position of particles, the columns of the descriptor matrix $\mathbf{X}$ correspond to the N vertices of a three dimensional polyhedron. Consequently, the three N-dimensional vectors, corresponding to the vectors gathering in turn the $\{x, y, z\}$ say, position space coordinate sets of all the particles, just describe a triangle in N-dimensional space.

However, not all the possible infinite N-dimensional triangles are *feasible*, that is: not all the triangles one can construct in description space can be attached to the set of N particles in object space. Any triangle belonging to the feasible set has to be associated to the following condition, here given as an algorithm, involving the squared Euclidian distances between all two object position vectors[3]:

$$\forall P, Q \in [1, \ldots N] \wedge P \neq Q : D_{PQ}^2 = \left\langle \left(|\mathbf{c}_P\rangle - \left|\mathbf{c}_Q\right\rangle \right) * \left(|\mathbf{c}_P\rangle - \left|\mathbf{c}_Q\right\rangle \right) \right\rangle = \langle |\mathbf{c}_P\rangle * |\mathbf{c}_P\rangle\rangle + \left\langle \left|\mathbf{c}_Q\right\rangle * \left|\mathbf{c}_Q\right\rangle \right\rangle - 2\left\langle |\mathbf{c}_P\rangle * \left|\mathbf{c}_Q\right\rangle \right\rangle > 0. \quad (4)$$

To be more precise, any feasible triangle made of the three N-dimensional descriptor vectors, has to adapt to the property shown in Eq. (4), which can be also alternatively and formally expressed as:

$$T = \{\langle \mathbf{f}_I | \, | I = 1, 3\} \in \Delta_F \rightarrow \forall P, Q \in [1, \ldots N] \wedge P \neq Q : \delta \left[D_{PQ}^2 > 0 \right], \quad (5)$$

where Δ_F is the set of feasible triangles, $\delta[L]$ is a logical Kronecker delta, which yields 0 if the logical sentence L is false and returns 1, whenever L is true.

[3] The symbol $*$ involving two vectors of a given vector space: $\forall |\mathbf{a}\rangle, |\mathbf{b}\rangle \in \mathrm{V}_D : |\mathbf{c}\rangle = |\mathbf{a}\rangle * |\mathbf{b}\rangle \rightarrow \forall I : c_I = a_I b_I \rightarrow |\mathbf{c}\rangle \in \mathrm{V}_D$ denotes an *inward* product, acting on two vectors and yielding another vector of the same vector space. The symbol involving a vector: $\alpha = \langle |\mathbf{a}\rangle \rangle = \sum_I a_I$ corresponds to the *complete* sum of its elements. Therefore, the expression: $\langle |\mathbf{a}\rangle * |\mathbf{b}\rangle \rangle = \sum_I a_I b_I \equiv \langle \mathbf{a}|\mathbf{b}\rangle$ corresponds to the scalar product of the two vectors. In the previous definitions have been used column vectors but the same definitions hold for row vectors, matrices or hypermatrices. If the involved vectors are functions: $|f\rangle \equiv f(\mathbf{r})$, the inward products $|f\rangle * |g\rangle \equiv f(\mathbf{r}) g(\mathbf{r})$ are coincident with products of functions, and the complete sum of a function becomes an integral over the definition domain: $\langle f(\mathbf{r}) \rangle = \int_D f(\mathbf{r}) d\mathbf{r}$. A scalar product of two functions can be written in this notation as: $\langle |f\rangle * |g\rangle \rangle = \int_D f(\mathbf{r}) g(\mathbf{r}) d\mathbf{r} \equiv \langle f|g\rangle$.

3.1.1 Origin shifts

Any triangle in a space of arbitrary dimension can be manipulated keeping invariant the three distances between its vertices, using in any order the three possible origin shifts:

$$\forall I, J, K \in (1, 2, 3) \wedge I \neq J \neq K : \left\langle {}^{I}\mathbf{g}_{J} \right| = \langle \mathbf{f}_{J} | - \langle \mathbf{f}_{I} | \wedge \left\langle {}^{I}\mathbf{g}_{K} \right| = \langle \mathbf{f}_{K} | - \langle \mathbf{f}_{I} | \, ; \tag{6}$$

such origin shift is the same as to transform the vector $\langle \mathbf{f}_{I} | \rightarrow \langle \mathbf{0} |$, thus it is also equivalent to consider one vertex of the triangle as the origin of coordinates, while constructing the two remnant triangle vertices with the shifted vectors $\left\{ \left\langle {}^{I}\mathbf{g}_{J} \right| ; \left\langle {}^{I}\mathbf{g}_{K} \right| \right\}$.

The resulting origin shifted triangle, provides the same information about the feasible triangle, but with only two linearly independent vectors needed instead of three. Such a property is common to linearly independent sets of vectors: the origin shift transform them into a set of cardinality lowered by one unit [3].

3.2 Molecular description problems

When the problem consists of a number of N molecular structures, each one attached to some M-dimensional vector of molecular descriptors, as commented before the resulting descriptor matrix $\mathbf{X}$ is of dimension $(M \times N)$. The N-dimensional row vectors describe a polyhedron of M vertices. As the set of distances between these vertices is made in general of different lengths, it can be called a N-dimensional *descriptor complex*. At the same time, the M-dimensional set made by the N vector columns, constitutes the vertices of an *object complex*, located within such a M-dimensional vector space.

In classical quantitative structure-properties relations (QSPR), the descriptor matrix is constructed as a first step to be used to find out linear functionals, connecting the descriptors with an appropriate set of optimal scalars to yield a model to estimate molecular properties. In QSPR, several basic algorithms performing this task are used, depending on statistical procedures; see for instance [10–21]. In order to avoid statistical overparameterization, some modeling algorithms project the descriptor space into an optimal parameter space of fewer dimensions: $m \ll M$. So, in this way, after such statistical manipulation the projected descriptor matrix becomes $(m \times N)$ dimensional.

However, in general such dimension reducing statistical procedures produce well studied problems; see for example references [13,14]. In addition, the so-called *dimensionality paradox* [1,2,4] also appears in classical QSPR problems within the descriptor space environment.

3.2.1 The dimensionality paradox

The origin of the dimensionality paradox might be associated to the manipulation of descriptor spaces. It can be connected with the fact that, starting with a set of

different molecules, an initial set of M-dimensional linearly *independent* descriptor vectors is built up, in order to represent the chosen molecular structures. Subsequent dimension reduction to m-dimensional vectors of the described molecular set, with the additional relationship: $m \ll N \ll M$, produces a newly described molecular set, consisting of N linearly *dependent* described molecules. Dimensionality paradox appears at this stage, since such a result contradicts the fact that the initially described molecular set is made of different linearly independent structures, logically attached to different molecules. Dimensionality paradox can be obviated studying any QSPR problem within the object space, the space of molecules instead of the descriptor space, for more details see [1,2].

The projected m row N-dimensional vectors of the classical QSPR manipulation form a complex with fewer vertices than the original setup, of course. When taking into account that just the top dimension 3 must be substituted by m, then the same properties as in Eqs. (4) and (5) can be transferred in the present framework, and Eq. (6) is equally valid. This is equivalent to say that not all complexes of the reduced number of m vertices are feasible at any representation level.

3.3 Résumé

In fact, after this preliminary discussion, one can realize that both position coordinate and description problems can be seen as particular cases of a unique formalism, where a set of N objects is described within an M-dimensional space by a set of chosen vectors.

This process produces a descriptor matrix of dimension $(M \times N)$ and in company of this mathematical construct, also appears a twofold dimensionality framework.

Two complexes made of M and N vertices respectively can be assembled, because of the inherent twofold dimensionality of all the problems connected with the nature of descriptor matrices.

4 Object origin shifts

Considering now again the column partition of the descriptor matrix, $C = \{|\mathbf{c}_J\rangle \,|J = 1, N\}$ as defined in Eq. (2), the associated N-vertex, M-dimensional object complex can be origin shifted in a similar way as when the descriptor complexes have been studied. In any case, then one of them can be chosen to perform an origin shift in the same way as in the previous position problems. It is a matter to define the shifted column vector vertex sets:

$$\forall I = 1, N : \mathrm{D}_I = \left\{\forall J = 1, N \wedge J \neq I : \left|{}^I\mathbf{d}_J\right\rangle = |\mathbf{c}_J\rangle - |\mathbf{c}_I\rangle\right\} \tag{7}$$

The set of shifted object complexes $\{\mathrm{D}_I \,|I = 1, N\}$ can be considered as several equivalent ways to observe the object set partition of the descriptor matrix. In every shifted set the distances between the complex vertices are invariant.

However, by origin shifting the angles subtended between two vertices are not invariant. To see this, consider two vertices $\{P, Q\}$ shifted with two different vertex shifts $\{R, S\}$ say. The corresponding cosine of the subtended angle in the shift by Rcan be written as:

$$\cos\left({}^{R}\alpha_{PQ}\right) = \left\langle \left|{}^{R}\mathbf{d}_P\right\rangle * \left|{}^{R}\mathbf{d}_Q\right\rangle \right\rangle \left(\left\langle \left|{}^{R}\mathbf{d}_P\right\rangle * \left|{}^{R}\mathbf{d}_P\right\rangle \right\rangle \left\langle \left|{}^{R}\mathbf{d}_Q\right\rangle * \left|{}^{R}\mathbf{d}_Q\right\rangle \right\rangle\right)^{-\frac{1}{2}}$$

while the corresponding cosine in the shift S might be written as:

$$\cos\left({}^{S}\alpha_{PQ}\right) = \left\langle \left|{}^{S}\mathbf{d}_P\right\rangle * \left|{}^{S}\mathbf{d}_Q\right\rangle \right\rangle \left(\left\langle \left|{}^{S}\mathbf{d}_P\right\rangle * \left|{}^{S}\mathbf{d}_P\right\rangle \right\rangle \left\langle \left|{}^{S}\mathbf{d}_Q\right\rangle * \left|{}^{S}\mathbf{d}_Q\right\rangle \right\rangle\right)^{-\frac{1}{2}}$$

It is a matter to develop the two leading scalar products, for instance:

$$\begin{aligned}\left\langle \left|{}^{R}\mathbf{d}_P\right\rangle * \left|{}^{R}\mathbf{d}_Q\right\rangle \right\rangle &= \left\langle (|\mathbf{c}_P\rangle - |\mathbf{c}_R\rangle) * \left(\left|\mathbf{c}_Q\right\rangle - |\mathbf{c}_R\rangle\right)\right\rangle \\ &= \left\langle |\mathbf{c}_P\rangle * \left|\mathbf{c}_Q\right\rangle \right\rangle + \langle |\mathbf{c}_R\rangle * |\mathbf{c}_R\rangle\rangle - \left\langle \left(|\mathbf{c}_P\rangle + \left|\mathbf{c}_Q\right\rangle\right) * |\mathbf{c}_R\rangle \right\rangle\end{aligned}$$

and

$$\begin{aligned}\left\langle \left|{}^{S}\mathbf{d}_P\right\rangle * \left|{}^{S}\mathbf{d}_Q\right\rangle \right\rangle &= \left\langle (|\mathbf{c}_P\rangle - |\mathbf{c}_S\rangle) * \left(\left|\mathbf{c}_Q\right\rangle - |\mathbf{c}_S\rangle\right)\right\rangle \\ &= \left\langle |\mathbf{c}_P\rangle * \left|\mathbf{c}_Q\right\rangle \right\rangle + \langle |\mathbf{c}_S\rangle * |\mathbf{c}_S\rangle\rangle - \left\langle \left(|\mathbf{c}_P\rangle + \left|\mathbf{c}_Q\right\rangle\right) * |\mathbf{c}_S\rangle \right\rangle\end{aligned}$$

and write its difference using the expression:

$$\begin{aligned}&\left\langle \left|{}^{R}\mathbf{d}_P\right\rangle * \left|{}^{R}\mathbf{d}_Q\right\rangle \right\rangle - \left\langle \left|{}^{S}\mathbf{d}_P\right\rangle * \left|{}^{S}\mathbf{d}_Q\right\rangle \right\rangle \\ &\quad = \langle |\mathbf{c}_R\rangle * |\mathbf{c}_R\rangle\rangle - \langle |\mathbf{c}_S\rangle * |\mathbf{c}_S\rangle\rangle + \left\langle \left(|\mathbf{c}_P\rangle + \left|\mathbf{c}_Q\right\rangle\right) * (|\mathbf{c}_S\rangle - |\mathbf{c}_R\rangle)\right\rangle\end{aligned}$$

which will be non-zero unless $|\mathbf{c}_R\rangle = |\mathbf{c}_S\rangle$. This property indicates that, while distances are preserved, the angles of the shifted complexes might be different.

The different shifted object complexes generate a set of N shifted descriptor matrices. Their construction can be written by means of the algorithm:

$$\forall I = 1, N : \mathbf{Y}_I = \left\{ \delta\left[P \neq I\right] \left|{}^{I}\mathbf{d}_P\right\rangle \,|\, P = 1, N \right\}$$

where the logical Kronecker delta assigns the vector zero to the I-th vector, that is: $\left|{}^{I}\mathbf{d}_I\right\rangle = |\mathbf{0}\rangle$.

The set of the shifted descriptor matrices considered as a whole generates a set of possible linearly independent elements. The linear dependence of the set of matrices: $\mathrm{Y} = \{\mathbf{Y}_I \,|\, I = 1, N\}$ can be tested via the scalar product Gram matrix:

$$\mathbf{\Gamma} = \{\Gamma_{IJ} = \langle \mathbf{Y}_I * \mathbf{Y}_J \rangle \,|I, J = 1, N\}.$$

The scalar products forming the Gram matrix can be obtained with the following algorithm:

$$\forall I, J : \langle \mathbf{Y}_I * \mathbf{Y}_J \rangle = \sum_P \delta\,[P \neq I \wedge P \neq J] \left\langle \left| {}^I\mathbf{d}_P \right\rangle * \left| {}^J\mathbf{d}_P \right\rangle \right\rangle.$$

5 Convex vector origin shifts and the particular case of centroid shift

The origin shifts discussed until now are attached to the raw vectors forming the rows or columns of descriptor matrices. However, these origin shifts are not the unique ones which can be performed on the resulting complexes. In fact, the problem has been discussed in another context [3,6], but here the nature of the object description merits another discussion for the sake of completeness.

When facing the N object M-dimensional description vector set: $C = \{|\mathbf{c}_J\rangle \,| \, J = 1, N\}$, then knowing a convex set of scalars, which can be defined as: $\mathrm{A} = \left\{\alpha_I \in \mathrm{R}^+ \,|I = 1, N\right\} \rightarrow \sum_I \alpha_I = 1$, a new vector can be constructed, such that: $|\mathbf{c}_\mathrm{A}\rangle = \sum_I \alpha_I \,|\mathbf{c}_I\rangle$.

In fact, convex sets as defined here can be associated to discrete probability distributions too. Among the infinite number of choices about the elements composing the convex set A, the specific uniform probability convex coefficient set, provides the *centroid* of the object complex, which can be written as: $|\mathbf{c}_C\rangle = N^{-1} \sum_I |\mathbf{c}_I\rangle$.

In general, the shifted object complex vertices by using any vertex convex combination can be described with the following set:

$$\mathrm{D_A} = \left\{\forall J = 1, N : \left| {}^\mathrm{A}\mathbf{d}_J \right\rangle = |\mathbf{c}_J\rangle - |\mathbf{c}_\mathrm{A}\rangle \right\},$$

every shifted vector can be expressed in turn via the following equalities:

$$\begin{aligned}
\forall J : \left| {}^\mathrm{A}\mathbf{d}_J \right\rangle &= |\mathbf{c}_J\rangle - |\mathbf{c}_\mathrm{A}\rangle = |\mathbf{c}_J\rangle - \sum_I \alpha_I \,|\mathbf{c}_I\rangle \\
&= (1 - \alpha_J)\,|\mathbf{c}_J\rangle - \sum_{I \neq J} \alpha_I \,|\mathbf{c}_I\rangle = \left(\sum_{I \neq J} \alpha_I \right) |\mathbf{c}_J\rangle - \sum_{I \neq J} \alpha_I \,|\mathbf{c}_I\rangle \\
&= \sum_{I \neq J} \alpha_I \,(|\mathbf{c}_I\rangle - |\mathbf{c}_J\rangle) = \sum_{I \neq J} \alpha_I \left| {}^J\mathbf{d}_I \right\rangle
\end{aligned}$$

yielding a final result which proves that the operation of shifting every J-th object complex vertex with respect any arbitrary convex linear combination of object vertices, becomes the same action as performing the same convex linear combination of the complex vertices shifted by the J-th vertex.

This property also provides the following and obvious centroid result, which can be taken as a particular case of the previous convex shifting:

$$\forall J : \left|{}^{\mathrm{C}}\mathbf{d}_J\right\rangle = |\mathbf{c}_J\rangle - |\mathbf{c}_C\rangle = N^{-1}\sum_{I\neq J}\left|{}^{J}\mathbf{d}_I\right\rangle$$

that is: the centroid shifting of the J-th object, belonging to the complex vertices, becomes the same as the arithmetic average of all the origin shifted vertices with respect the J-th vertex.

Because of the descriptor matrix twofold dimensionality problem, the properties obtained concerning convex combinations of the object complex vertices are totally reproducible over the descriptor complex vertices. To do this two way exchange of properties, it is just needed to change columns by rows and to perform the appropriate changes of dimensions and number of complex vertices.

6 Rotations of descriptor matrices

Origin shifts can be considered well-defined translations within the twofold dimension spaces of the descriptor matrices. The geometrical point of view which has been adopted until now can also lead to study rotations of the elements of such matrices. Rotations of descriptor matrices can be performed over their twofold dimensions.

As it is well-known, rotations in vector spaces of arbitrary dimension, defined over the real field, as the proposed problems are, correspond to multiplication by orthogonal matrices $\mathbf{U} = \{U_{IJ}\}$ of the appropriate dimension, which in general, besides to be square and non-singular, possess as the main property: $\mathbf{U}\mathbf{U}^T = \mathbf{U}^T\mathbf{U} = \mathbf{I}$; that is, the transpose matrix: $\mathbf{U}^T = \{U_{IJ}^{(T)} = U_{JI}\}$ has to be coincident with the inverse: $\mathbf{U}^T = \mathbf{U}^{-1}$. Therefore, the orthogonal matrices determinant is necessarily associated to the property: $Det\,|\mathbf{U}| = \pm 1$.

Rotations of any $(M \times N)$ descriptor matrix can be easily performed by multiplying on the left by an orthogonal $(M \times M)$ matrix: ${}_M\mathbf{U}$ and on the right by an $(N \times N)$ orthogonal matrix: ${}_N\mathbf{U}$: $({}_M\mathbf{U})\,\mathbf{X}\,({}_N\mathbf{U}) \rightarrow \mathbf{X}_U$. The resulting rotated matrix $\mathbf{X}_U$ has the twofold dimension invariant. Of course, rotations might be performed just on the left or on the right only, besides of both sides as earlier shown.

Left rotations act on descriptor space vectors, while left rotations act on object space vectors. Rotations leave invariant both distances and angles of the corresponding complexes. However, suppose a rotation is performed over the descriptor complex by an orthogonal $(M \times M)$ matrix $\mathbf{V}$, named in this way in order to simplify the previously used notation, that is:

$$\mathbf{V}\mathbf{X} = \mathbf{Y} \rightarrow \forall I, J : Y_{IJ} = \sum_K V_{IK}X_{KJ}.$$

This is equivalent to transform the column partition of the descriptor matrix in the following way:

$$\forall J = 1, N : \mathbf{V}\,|\mathbf{c}_J\rangle = |\mathbf{d}_J\rangle \rightarrow \mathbf{Y} = (|\mathbf{d}_1\rangle\,;|\mathbf{d}_2\rangle \ldots |\mathbf{d}_N\rangle)\,.$$

Scalar products within the rotated column vectors remain invariant, as it is well-known; they can be written:

$$\forall P, Q \in [1, \ldots N] : \left\langle |\mathbf{d}_P\rangle * \left|\mathbf{d}_Q\right\rangle\right\rangle = \sum_I d_{IP} d_{IQ} = \sum_I \sum_K \sum_L V_{IK} V_{IL} c_{KP} c_{LQ}$$
$$= \sum_K \sum_L \left(\sum_I V_{LI}^{(T)} V_{IK} \right) c_{KP} c_{LQ} = \sum_K \sum_L \delta_{LK} c_{KP} c_{LQ}$$
$$= \sum_K c_{KP} c_{KQ} = \left\langle |\mathbf{c}_P\rangle * \left|\mathbf{c}_Q\right\rangle\right\rangle$$

Therefore the distances and angles subtended by the column spaces of the descriptor matrices $\mathbf{X}$ and $\mathbf{Y}$ are invariant. However, once performed the rotation, the resultant structure of the row space of the rotated descriptor matrix can be written easily as:

$$\mathbf{Y} = \begin{pmatrix} \langle \mathbf{g}_1 | \\ \langle \mathbf{g}_2 | \\ \vdots \\ \langle \mathbf{g}_M | \end{pmatrix} \rightarrow \forall I = 1, M : \langle \mathbf{g}_I | = \left\{ Y_{IJ} = \sum_K V_{IK} X_{KJ} \,|J = 1, N \right\}$$

consequently the scalar products of the resultant rows can be written:

$$\forall P, Q \in [1, \ldots M] : \left\langle \langle \mathbf{g}_P | * \left\langle \mathbf{g}_Q \right| \right\rangle = \sum_S \sum_K \sum_L V_{PK} V_{QL} X_{KS} X_{LS}$$
$$= \sum_K \sum_L V_{PK} V_{QL} \, \langle \langle \mathbf{f}_K | * \langle \mathbf{f}_L | \rangle$$

yielding an expression, which is no longer invariant. Thus, when performing rotations on the object complex the descriptor complex is deformed according to the chosen rotation.

For instance, when rotating the coordinate vectors of a position problem, the relative positions of the described particles in three dimensional spaces are preserved, but the triangle of the position coordinates vectors in N-dimensional space might be changed accordingly. Of course, this is reversed when performing rotations on the descriptor space, which permit the invariance of the corresponding complex, but the object complex is no longer invariant, becoming deformed with respect the initial structure.

7 Gram matrices generated in object and descriptor spaces

Once a descriptor matrix is known, one can manipulate it, according to the twofold final dimension of the wanted resultant space. For instance, in QSPR one can obtain two kinds of square Gram matrices. One can be generated with a dimension $(M \times M)$

which corresponds to the descriptor space:

$$\mathbf{\Gamma}_D = \mathbf{X}\mathbf{X}^T \rightarrow \Gamma_{D;IJ} = \langle \langle \mathbf{f}_I | * \langle \mathbf{f}_J | \rangle = \sum_K X_{IK} X_{JK} \equiv \langle \mathbf{f}_I | \mathbf{f}_J \rangle \tag{8}$$

While the other can be of dimension $(N \times N)$ and might be obtained via the product:

$$\mathbf{\Gamma}_O = \mathbf{X}^T\mathbf{X} \rightarrow \Gamma_{O;IJ} = \langle | \mathbf{c}_I \rangle * | \mathbf{c}_J \rangle \rangle = \sum_K X_{KI} X_{KJ} \equiv \langle \mathbf{c}_I | \mathbf{c}_J \rangle \tag{9}$$

and corresponds to the object space.

The matrix defined in Eq. (8) is the basic starting point of classical QSPR procedures [10–21], while the matrix defined in Eq. (9) is promoted by quantum QSPR procedures, as a mean to avoid the dimensionality paradox and obtain similar results as in classical QSPR, see for example [1,2].

7.1 The three dimensional position cases

It is to be noted the structure of the matrix described in Eq. (8) when the descriptor space is of dimension three. Obviously enough, the resultant matrix will be (3×3) in the cases of particle description in three-dimensional spaces. To simplify the notation the descriptor matrix and the corresponding transpose can be written by means of the column matrix:

$$\mathbf{X} = \begin{pmatrix} \langle \mathbf{x}_1 | \\ \langle \mathbf{x}_2 | \\ \langle \mathbf{x}_3 | \end{pmatrix} \wedge \mathbf{X}^T = \begin{pmatrix} | \mathbf{x}_1 \rangle & | \mathbf{x}_2 \rangle & | \mathbf{x}_3 \rangle \end{pmatrix}$$
$$\rightarrow \mathbf{\Gamma}_D = \left\{ \Gamma_{D;IJ} = \langle \mathbf{x}_I | \mathbf{x}_J \rangle \equiv \langle | \mathbf{x}_I \rangle * | \mathbf{x}_J \rangle \rangle \, | I, J = 1, 3 \right\}$$

where the three row vectors $\{\langle \mathbf{x}_I | \, | I = 1, 3\}$ contain the respective particle $\{x, y, z\}$ coordinates.

Thus defined, the elements of the symmetric matrix $\mathbf{\Gamma}_D$ contains scalar products of the all pairs of the whole set of coordinates. In fact, such matrix is closely related to the tensor, which appears in the background framework of both the moment of inertia and quadrupole tensors of a many particle system. Therefore, as the previously mentioned tensors are respectively weighted by masses and charges, the matrix $\mathbf{\Gamma}_D$ might be considered as an *unweighted* characteristic of the set, representing every precise situation of the many particles three dimensional positioning.

Principal axis and components, what is the same eigenvectors and eigenvalues, of $\mathbf{\Gamma}_D$ could be valuable to describe schematically the spatial distribution of particle swarms. In fact, the matrix $\mathbf{\Gamma}_D$ describes both moment of inertia and quadrupole moment of an uniform particle set, where each particle element possess the same mass and the same charge, like an electron, atomic gas or plasma made of the same particles. It can be called *characteristic form* tensor and attached to any particle swarm or molecular conformation, despite it is not weighted.

7.2 The molecular conformation scenarios

In molecular conformations particle systems positioning, all linear molecules will possess two null eigenvalues of the characteristic molecular form tensor, while planar molecules will have one null eigenvalue. Three degenerate eigenvalues preclude spherical molecules, while two degenerate eigenvalues will denote elongated molecular structures in one space direction. The eigenvalue set of the molecular characteristic form tensor could be a good résumé of the attached molecular structure and might be employed as a molecular descriptor, as similar other molecular parameters of this kind have been employed [26].

Even one can use the molecular characteristic form tensor to obtain simple differentiation discrete descriptors, permitting to construct optical isomers schematic numerical sketches. As one can describe in this case two conformation descriptor matrices: $\{\mathbf{X}_R; \mathbf{X}_S\}$; then, the eigenvalues of both $\mathbf{\Gamma}_D$ matrices:

$$^{R}\mathbf{\Gamma}_D = \mathbf{X}_R\mathbf{X}_R^T \wedge ^{S}\mathbf{\Gamma}_D = \mathbf{X}_S\mathbf{X}_S^T \rightarrow ^{R}\mathbf{\Gamma}_D = ^{S}\mathbf{\Gamma}_D = \mathbf{G}_D$$

will be the coincident, but different from the hybrid matrices:

$$^{RS}\mathbf{\Gamma}_D = \mathbf{X}_R\mathbf{X}_S^T \wedge ^{SR}\mathbf{\Gamma}_D = \mathbf{X}_S\mathbf{X}_R^T \rightarrow ^{SR}\mathbf{\Gamma}_D = \left(^{RS}\mathbf{\Gamma}_D\right)^T.$$

Therefore it might be worthwhile to consider the eigensystem of the symmetric matrix:

$$\mathbf{A}_{RS} = \mathbf{X}_R\mathbf{X}_S^T + \mathbf{X}_S\mathbf{X}_R^T$$

when compared with the pure R or S characteristic molecular form tensors; that is: the eigensystem of the difference matrix:

$$\mathbf{\Delta}_{RS} = \frac{1}{2}\left(\mathbf{X}_R\mathbf{X}_R^T + \mathbf{X}_S\mathbf{X}_S^T\right) - \frac{1}{2}\left(\mathbf{X}_R\mathbf{X}_S^T + \mathbf{X}_S\mathbf{X}_R^T\right) = \mathbf{G}_D - \frac{1}{2}\mathbf{A}_{RS}$$

It is plausible to expect that the $\mathbf{\Delta}_{RS}$ eigensystem can be a good measure of the structural difference between both isomers.

A recent study has proposed the definition of a *plane of best fit* [22], as a way to characterize and quantify the three dimensional character of molecular structures, alternative to the ones reported by other authors [23–25] and of these summarized by Todeschini and Consonni [26].

In any case, one can be confident that the molecular characteristic form tensor obtained as a consequence of a more general framework involving descriptor matrices, will provide a set of logical parameters to tackle with molecular shape in three dimensional spaces. Moreover, the related quadrupole moment tensor principal components have been used several years ago [27] and nowadays [28] in our laboratory as an initial positioning to orient in a general way coordinate axis of molecular structures. Therefore, the simplified configuration provided by the molecular characteristic form tensor might be of interest for these problems.

7.3 Quantum similarity matrices

The Gram matrix over the object space is equivalent to the quantum similarity matrix [29,30] constructed from the density function tag set of any quantum object set [31, 32]. This connection between both matrix structures corresponds to the following construction quantum similarity algorithm.

Starting form a set of density functions associated to a set of well-defined quantum objects, usually molecules: $\mathrm{P} = \{\rho_I\,(\mathbf{r})\,|I = 1, N\}$, then the row vector which contains as elements the densities of P can be constructed as:

$$|\mathbf{X}\rangle = \left(\rho_1\ \rho_2 \cdots \rho_N\right) \equiv \left(|\rho_1\rangle\ |\rho_2\rangle \cdots |\rho_N\rangle\right) \tag{10}$$

A matrix like (10) can be named as *continuous* descriptor matrix. The vector expressed in Eq. (10), is the equivalent to a discrete descriptor matrix, as the one presented at the beginning in Eqs. (1) and (2), with the left row dimension becoming infinite though. That is, one can express the twofold dimension corresponding to this situation as: $(\infty \times N)$. Thus the equivalent to the object Gram matrix (9) can be expressed formally by means of the so called similarity matrix[4]:

$$\mathbf{\Gamma}_O = [\![|\mathbf{X}\rangle \otimes |\mathbf{X}\rangle]\!] \equiv [\![|\mathbf{X}\rangle\,\langle\mathbf{X}|]\!] \rightarrow \Gamma_{O;IJ} = Z_{IJ} = \langle\rho_I\rho_J\rangle = \int_D \rho_I\,(\mathbf{r})\,\rho_J\,(\mathbf{r})\,d\mathbf{r}$$

where in addition it has been used the habitual symbol for the similarity matrix and its elements: $\mathbf{Z} = \{Z_{IJ}\}$. Therefore, there is a clear connection between quantum similarity matrices and Gram matrices within object spaces in discrete descriptor matrices.

8 Discussion and results

Straightforward analysis of real $(M \times N)$ descriptor matrices, connected with theoretical chemical and physical problems reveals several interesting features. The discussion of these common elements permits to employ a geometrical point of view to disclose properties, transformation characteristics and application possibilities common to any problem which can be associated to descriptor matrices. The proposed insight can be useful from Monte Carlo statistical mechanics up to QSPR.

References

1. R. Carbó-Dorca, A. Gallegos, Á.J. Sánchez, Notes on quantitative structure-properties relationships (QSPR) (1): a discussion on a QSPR dimensionality paradox (QSPR DP) and its quantum resolution. J. Comput. Chem. **30**, 1146–1159 (2008)

[4] Provided that the symbol $[\![\mathbf{A}]\!]$, applied over a matrix with functions as elements: $\mathbf{A}\,(\mathbf{r}) = \{A_{IJ}\,(\mathbf{r})\}$, might be considered that yields the matrix of the integrals of the function elements: $\mathbf{G} = [\![\mathbf{A}]\!] = \left\{G_{IJ} = \int_D A_{IJ}\,(\mathbf{r})\,d\mathbf{r}\right\}$.

2. R. Carbó-Dorca, Notes on quantitative structure-properties relationships (QSPR) (3): density functions origin shift as a source of quantum QSPR (QQSPR) algorithms in molecular spaces. J. Comput. Chem. (2012). doi:10.1002/jcc.23198.
3. R. Carbó-Dorca, E. Besalú, Centroid origin shift of quantum object sets and molecular point clouds: description and element comparisons. J. Math. Chem. **50**, 1161–1178 (2012)
4. R. Carbó-Dorca, Quantum similarity, in *Concepts and Methods in Modern Theoretical Chemistry*, ed. by S.K. Ghosh, P.K. Chattaraj, vol. 1 (Taylor & Francis, London, 2013)
5. R. Carbó-Dorca, Mathematical aspects of the LCAO MO first order density function (5): centroid shifting of MO ShF basis set, properties and applications. J. Math. Chem. (2012). doi:10.1007/s10910-012-0083-x
6. R. Carbó-Dorca, A naïve geometrical perspective of Fukui functions: definition of Fukui function skew symmetric matrices described on density function sets. J. Math. Chem. (2012). doi:10.1007/s10910-012-0120-9
7. R. Carbó-Dorca, E. Besalú, Shells, point cloud huts, generalized scalar products, cosines and similarity tensor representations in vector semispaces. J. Math. Chem. **50**, 210–219 (2012)
8. R. Carbó-Dorca, About the concept of chemical space: a concerned reflection on some trends of modern scientific thought within theoretical chemical lore. J. Math. Chem. (2012). doi:10.1007/s10910-012-0091-x
9. R. Carbó-Dorca, Enfolded conformational spaces: definition of the chemical quantum mechanical multiverse under Born–Oppenheimer approximation. IQC technical report TR-2012-12. J. Math. Chem. (submitted)
10. P. Bultinck, H. De Winter, W. Langenaeker, J.P. Tollenaere (eds.), *Molecular Similarity and QSAR in Computational Medicinal Chemistry for Drug Discovery* (Marcel Dekker Inc., New York, 2004)
11. J.P. Doucet, A. Panaye, *Three Dimensional QSAR* (CRC Press, Boca Raton, FL, 2010)
12. Y.C. Martin, *Quantitative Drug Design: A Critical Introduction* (CRC Press, Boca Raton, FL, 2010)
13. J.C. Dearden, M.T.D. Cronin, K.L.E. Kaiser, SAR QSAR Environ. Res. **20**, 241 (2009)
14. M.T.D. Cronin, T.W.J. Schultz, Mol. Struct. (Theochem) **622**, 39 (2003)
15. W. Karcher, J. Devilliers (eds.), *Practical Applications of QSAR in Environmental Chemistry and Toxicology* (Kluwer, Dordrecht, 1990)
16. H. Kubinyi (ed.), *3D QSAR in Drug Design* (ESCOM, Leiden, 1993)
17. C. Hansch, C. Leo, *Exploring QSAR. ACS Professional Reference Book* (ACS, Washington DC, 1995)
18. F. Sanz, J. Giraldo, F. Manault (eds.), *QSAR and Molecular Modelling* (Prous Science, Barcelona, 1995)
19. N.C. Cohen (ed.), *Molecular Modelling in Drug Design* (Academic Press, San Diego, CA, 1996)
20. J. Devilliers (ed.), *Comparative QSAR* (CRC Press, Boca Raton, FL, 1998)
21. R. Carbó-Dorca, D. Robert, L. Amat, X. Gironés, E. Besalú, Molecular quantum similarity in QSAR and drug design, in *Lecture Notes in Chemistry*, vol. 73 (Springer, Berlin, 2000).
22. N.C. Firth, N. Brown, J. Blagg, Plane of best fit: a novel method to characterize the 3D of molecules. J. Chem. Inf. Mod. **52**, 2516–2525 (2012)
23. W.H.B. Sauer, M.K. Schwarz, Molecular shape diversity of combinatorial libraries: a prerequisite for broad bioactivity. J. Chem. Inf. Comput. Sci. **43**, 987–1003 (2003)
24. F. Lovering, J. Bikker, C. Humblet, Escape from flatland: increasing saturation as an approach to improving clinical success. J. Med. Chem. **52**, 6752–6756 (2009)
25. A.Y. Meyer, Molecular mechanics and molecular shape. III. Surface area and cross-sectional areas of organic molecules. J. Comput. Chem. **7**, 144–152 (1986)
26. R. Todeschini, V. Consonni, *Molecular Descriptors for Chemoinformatics* (Wiley-VCH, Germany, 2009)
27. R. Carbó, B. Calabuig, Molsimil-88: molecular similarity calculations using a CNDO approximation. Comput. Phys. Commun. **55**, 117 (1989)
28. R. Carbó-Dorca, E. Besalú, L.D. Mercado, Communications on quantum similarity (3): a geometric-quantum similarity molecular superposition (GQSMS) algorithm. J. Comp. Chem. **32**, 582–599 (2011)
29. R. Carbó-Dorca, E. Besalú, A general survey of molecular quantum similarity Huzinaga symposium. Fukuoka. J. Molec. Struct. Theochem. **451**, 11–23 (1998)
30. R. Carbó-Dorca, L. Amat, E. Besalú, X. Gironès, D. Robert, *Quantum Molecular Similarity: Theory and Applications to the Evaluation of Molecular Properties, Biological Activity and Toxicity. Mathematical and Computational Chemistry: Fundamentals of Molecuar Similarity* (Kluwer Academic/Plenum Publishers, Dordrecht, 2001)

31. P. Bultinck, X. Gironés, R: Carbó-Dorca. Molecular quantum similarity: theory and applications. in, *Reviews in Computational Chemistry*, vol. 21, ed. by K.B. Lipkowitz, R. Larter and T. Cundari (Wiley, Hoboken, 2005), pp. 127–207
32. R. Carbó, B. Calabuig, Molecular quantum similarity measures and N-dimensional representation of quantum objects. I. Theoretical foundations. Int. J. Quantum Chem. **42**, 1681 (1992)