J Math Chem (2012) 50:734–740
DOI 10.1007/s10910-011-9920-6

ORIGINAL PAPER

Scaled Euclidian distances: a general dissimilarity index with a suitably defined geometrical foundation

Ramon Carbó-Dorca

Received: 16 February 2011 / Accepted: 23 September 2011 / Published online: 7 October 2011

Abstract A simple manipulation of the Euclidian distance expression permits to obtain a scaled dissimilarity index measure, varying within a range of values lying in the interval [0,1]. Here is presented the theoretical background, its application to quantum similarity and use for artificial intelligence general purposes as well. The origin of Hodgkin-Richards index is analyzed and compared with quantum similarity Carbó index.

Keywords Euclidian distances · Quantum similarity · Hodgkin-Richards similarity index · Carbó similarity index · Scaled Euclidian distance dissimilarity index · Generalized scalar products · Generalized Carbó similarity indices

1 Introduction

Since the dawn of quantum similarity [1], two indices: a similarity (SI) and a dissimilarity one (DI) were described. That is: (a) a cosine like index, which is now known as Carbó SI and (b) an Euclidian distance DI, which has remained so far unnamed (perhaps it can be called Carbó DI). With time both SI and DI have been studied from several points of view [2–4] and compared with other SI, which like the Hodgkin and Richards SI [5] have appeared in the literature. The Carbó SI has been studied and generalized in many papers [1–4,6,7], even at the present times, where some recent studies have been published [8–10] to deepen the understanding of this tool, as well as to smoothing its applicability [11–15].

On the other hand, the Euclidian distance DI has not been subject to so many studies as the Carbó SI and has been scarcely employed in quantum similarity applications,

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

see for example [16]. The reason of this difference between both indices has to be found in the convenient range of values, the easily generalization for the comparison of an indefinite number of quantum objects and the invariance properties of the Carbó SI.

For the time being it is difficult if not impossible to associate the same properties to the Euclidian distance DI, which by definition remains a two quantum object dissimilarity measure, based into the Euclidian norm of their density functions difference.

A good argument about the origin of this substantial difference between both kinds of indices can be found in the fact that the solid angle involving various quantum objects can be imagined in the appropriate dimensions. However, the definition of a unique distance between several quantum objects appears elusive. As far as we know, only can be easily imagined the construction of a zero diagonal matrix of the Euclidian distances between the involved object pairs.

In the present study it will be presented a trivial manipulation of the Euclidian distance DI, which permits a convenient scaling of the resultant computational values, in such a way that they remain in the interval [0,1]. Moreover, the connection of this scaled distance with the Hodgkin-Richards SI will be made evident.

2 Vector spaces, scalar products and Euclidian distances

Although distances in general have been many years ago the basic tool in taxonomy [17] and have been the subject of a modern exhaustive study [18], as it has been commented at the introduction, Euclidian distances have not yet been employed as DI in quantum similarity as has been the case with SI. To the aforementioned characteristics, one of the most annoying features of Euclidian distances is the fact they can have values in the interval $[0, +\infty]$ and this kind of large interval is certainly the reason of the Carbó SI most frequent choice. For this motive here will be discussed the possibility to construct a scaled Euclidian distance with more convenient features, having values varying in the interval [0,1] as Carbó SI, although preserving the DI properties. While this study is essentially adequate to the quantum similarity computational structure, it will be presented in a general vectorial space framework, so in this way the presented results can be applied on any problem described by real discrete or continuous vectors. For instance, involving the vector elements constructed in the classical way and intended to be used as discrete descriptors, as it is customary in QSAR studies in order to numerically represent molecular structures within N-dimensional spaces.

Taking into account normed metric vector spaces, defined in the real field $\mathbf{R}$, no matter which dimension possess, one can define a scalar product of two vectors in the usual way, but for generalization sake and further use, the following notation, see references [6] for more details, will be employed:

$$\forall \, |\mathbf{a}\rangle \, , \, |\mathbf{b}\rangle \in \mathrm{V}\,(\mathbf{R}) : \langle \mathbf{a}|\mathbf{b}\rangle = \langle |\mathbf{a}\rangle * |\mathbf{b}\rangle\rangle \in \mathbf{R}$$

where Dirac's notation has been chosen for vectors and the scalar product has been alternatively and formally defined with the aid of two operations:

a) The complete sum of a vector:

$$\forall \left|\mathbf{a}\right\rangle \in \mathrm{V}\left(\mathbf{R}\right) : \left\langle \left|\mathbf{a}\right\rangle \right\rangle \in \mathbf{R},$$

which corresponds to a linear operator acting on the reference vector space, and

b) The inward vector product:

$$\forall \left|\mathbf{a}\right\rangle, \left|\mathbf{b}\right\rangle \in \mathrm{V}\left(\mathbf{R}\right) : \left|\mathbf{a}\right\rangle * \left|\mathbf{b}\right\rangle \in \mathrm{V}\left(\mathbf{R}\right),$$

which produces another vector of the same kind as the factors entering the product; see for more details reference [15].

Using this notation, the squared Euclidian distance between two vectors can be written in the following way:

$$D_{ab}^2 = \left\langle \left(\left|\mathbf{a}\right\rangle - \left|\mathbf{b}\right\rangle\right) * \left(\left|\mathbf{a}\right\rangle - \left|\mathbf{b}\right\rangle\right) \right\rangle = \left\langle \mathbf{a}|\mathbf{a}\right\rangle + \left\langle \mathbf{b}|\mathbf{b}\right\rangle - 2\left\langle \mathbf{a}|\mathbf{b}\right\rangle. \tag{1}$$

Thus, Euclidian distances, are related to the symmetric (2×2) Gram matrix of the scalar products between the involved vectors. Taking into account that in $\mathrm{V}\left(\mathbf{R}\right)$ spaces the scalar product is commutative, so: $\left\langle \mathbf{b}|\mathbf{a}\right\rangle = \left\langle \mathbf{a}|\mathbf{b}\right\rangle$, then one can write such a matrix as:

$$\mathbf{Z} = \begin{pmatrix} \left\langle \mathbf{a}|\mathbf{a}\right\rangle & \left\langle \mathbf{a}|\mathbf{b}\right\rangle \\ \left\langle \mathbf{a}|\mathbf{b}\right\rangle & \left\langle \mathbf{b}|\mathbf{b}\right\rangle \end{pmatrix} = \begin{pmatrix} z_{aa} & z_{ab} \\ z_{ab} & z_{bb} \end{pmatrix} \tag{2}$$

and in this way the squared distance (1) can be expressed in a more compact manner as:

$$D_{ab}^2 = z_{aa} + z_{bb} - 2z_{ab}. \tag{3}$$

It is interesting to note that the Gram matrix (2) can be rewritten with little effort from the matrix $\mathbf{Z}_+$ forming one of the elements of the binary composition:

$$\mathbf{Z}_{\pm} = \begin{pmatrix} z_{aa} & 0 \\ 0 & z_{bb} \end{pmatrix} \pm \begin{pmatrix} 0 & z_{ab} \\ z_{ab} & 0 \end{pmatrix}$$

while Eq. (3) can be easily deduced by using the complete sum of the matrix elements of $\mathbf{Z}_-$:

$$D_{ab}^2 = \left\langle \mathbf{Z}_- \right\rangle.$$

Equation (3) constituting a classical formulation, which can be also easily rewritten like:

$$D_{ab}^2 = \left(z_{aa} + z_{bb}\right)\left(1 - \frac{z_{ab}}{\frac{1}{2}\left(z_{aa} + z_{bb}\right)}\right)$$

and calling half the trace of the Gram matrix:$s_{ab} = \frac{1}{2}(z_{aa} + z_{bb})$, one can arrive for the squared distance expression to propose a compact, not so usual form:

$$D_{ab}^2 = 2s_{ab}\left(1 - \frac{z_{ab}}{s_{ab}}\right). \tag{4}$$

In Eq. (4), the ratio:

$$h_{ab} = \frac{z_{ab}}{s_{ab}} \tag{5}$$

corresponds to an alternative to Carbó SI, proposed several years ago by Hodgkin and Richards [5].

In fact, due that the squared distance is definite non-negative, it is easily obtained the following set of characteristic properties:

$$\begin{array}{l} D_{ab}^2 = z_{aa} + z_{bb} - 2z_{ab} \geq 0 \rightarrow \\ z_{aa} + z_{bb} \geq 2z_{ab} \rightarrow s_{ab} \geq z_{ab} \rightarrow h_{ab} \in [0,1]\,. \end{array} \tag{6}$$

3 Carbó SI comparison

Instead of the Hodgkin and Richards arithmetic mean of the involved vector norms as a denominator in Eq. (5), the Carbó SI uses instead the geometric mean for this purpose:

$$r_{ab} = \frac{z_{ab}}{g_{ab}} \leftarrow g_{ab} = \sqrt{z_{aa}z_{bb}}.$$

As earlier commented, r_{ab} represents the cosine of the angle subtended by the two involved vectors in any metric space.

Also, a similar result as the one found in Eq. (6) comes from the determinant of the Gram matrix $\mathbf{Z}$ of Eq. (2) and the Carbó SI:

$$Det\,|\mathbf{Z}| = z_{aa}z_{bb} - z_{ab}^2 \geq 0 \rightarrow \sqrt{z_{aa}z_{bb}} \geq z_{ab} \rightarrow r_{ab} \in [0,1]\,.$$

At the same time, as the involved means of the two SI's: $h_{ab} \wedge r_{ab}$ are connected by the relationship: $s_{ab} \geq g_{ab}$, it will happen that both SI's are interrelated by the inequality: $r_{ab} \geq h_{ab}$.

3.1 Homothetic invariance of Carbó SI

Scaling the involved vectors in an independent homothetic way will produce, for example:

$$|\mathbf{a}\rangle \rightarrow \alpha\,|\mathbf{a}\rangle = |\mathbf{A}\rangle \wedge |\mathbf{b}\rangle \rightarrow \beta\,|\mathbf{b}\rangle = |\mathbf{B}\rangle\,. \tag{7}$$

Therefore, the Carbó SI will remain invariant as shows the following reasoning:

$$r_{AB} = \frac{\langle \mathbf{A}|\mathbf{B}\rangle}{\sqrt{\langle \mathbf{A}|\mathbf{A}\rangle \langle \mathbf{B}|\mathbf{B}\rangle}} = \frac{\alpha\beta \langle \mathbf{a}|\mathbf{b}\rangle}{\sqrt{\alpha^2 \langle \mathbf{a}|\mathbf{a}\rangle \beta^2 \langle \mathbf{b}|\mathbf{b}\rangle}} = r_{ab};$$

this property can be seen as a consequence of the fact that the Carbó SI corresponds to the cosine of the angle subtended by the two involved vectors and with the homothecy (7) the angle of both vectors do not vary through the scaling.

However, the Hodgkin and Richards SI behaves as:

$$h_{AB} = \frac{2\langle \mathbf{A}|\mathbf{B}\rangle}{(\langle \mathbf{A}|\mathbf{A}\rangle + \langle \mathbf{B}|\mathbf{B}\rangle)} = \frac{2\alpha\beta \langle \mathbf{a}|\mathbf{b}\rangle}{\left(\alpha^2 \langle \mathbf{a}|\mathbf{a}\rangle + \beta^2 \langle \mathbf{b}|\mathbf{b}\rangle\right)} = \frac{2\langle \mathbf{a}|\mathbf{b}\rangle}{\left(\frac{\alpha}{\beta} \langle \mathbf{a}|\mathbf{a}\rangle + \frac{\beta}{\alpha} \langle \mathbf{b}|\mathbf{b}\rangle\right)} \neq h_{ab},$$

it must be said that in this case both SI's become equal whenever both vectors have been multiplied by the same scalar factor as in a homogenous global homothecy. This property reminds of the Euclidian distance properties as:

$$D^2_{AB} = z_{AA} + z_{BB} - 2z_{AB} = \alpha^2 z_{aa} + \beta^2 z_{bb} - 2\alpha\beta z_{ab} \neq D^2_{ab}.$$

Unless it holds $\alpha = \beta$, a situation where, as it is well known, distances become scaled by the homothetic parameter:

$$D^2_{AB} = \alpha^2 D^2_{ab}.$$

3.2 Generalization of Carbó SI

The differences between both Hodgkin and Richards and Carbó SI's do not end with this homothetic variance-invariance feature, but with the generalization power embedded in the Carbó SI, in front of the Hodgkin and Richards SI.

Indeed, one can define a triple vector scalar product by means of the following algorithm, see references [7,11] for more details:

$$\forall\, |\mathbf{a}\rangle, |\mathbf{b}\rangle, |\mathbf{c}\rangle \in \mathrm{V}(\mathbf{R}) : \langle \mathbf{abc}\rangle = \langle |\mathbf{a}\rangle * |\mathbf{b}\rangle * |\mathbf{c}\rangle\rangle \in \mathbf{R},$$

in this manner, a triple vector Carbó SI can be also easily defined by means of:

$$r_{abc} = \frac{\langle \mathbf{abc}\rangle}{\sqrt[3]{\langle \mathbf{aaa}\rangle \langle \mathbf{bbb}\rangle \langle \mathbf{ccc}\rangle}}.$$

Provided that one redefines the triple products of the same vector as positive definite expressions; for instance, in the most simple and direct way, by means of taking absolute values of the result: $\langle \mathbf{aaa}\rangle \equiv |\langle \mathbf{aaa}\rangle|$. This problem doesn't appear when dealing with quantum object description through electronic density functions though. Being density functions non-negative defined, then the triple density scalar products of the

same density function are always positive definite. In fact, generally speaking the triple (or multiple) scalar products of density functions are also positive definite.

The extension of Carbó SI to quadruple and higher vector comparison algorithms can be found both in previous studies, see reference [7] for instance, and recent work [11,13–15] as well.

Such a Carbó SI generalization possibility cannot be performed in a simple way within the framework of Hodgkin and Richards SI. Thus, this limitation circumscribes a SI of this kind within the toolbox containing two vector comparative devices only. Such a situation is not surprising, taking into account that Hodgkin and Richards SI is naturally produced as a leading term in the manipulated expression of a squared Euclidian distance DI.

4 Scaled Euclidian distance characteristics

Moreover, returning to the squared distance expression (4) given in terms of the Hodgkin and Richards SI (5), one can deduce the Euclidian distance DI between both vectors by simply obtaining the square root, that is:

$$D_{ab} = \left(\sqrt{2s_{ab}}\right)\sqrt{(1 - h_{ab})}$$

an expression which can be rearranged, just taking away from the expression the trace of the Gram matrix **Z** in Eq. (2), yielding the following scaled Euclidian distance:

$$d_{ab} = \frac{D_{ab}}{\sqrt{2s_{ab}}} = \sqrt{(1 - h_{ab})}. \tag{8}$$

Now, the above scaled Euclidian distance of Eq. (8), corresponds to a DI, but with the additional feature of having values belonging to the interval [0,1]. This DI or its square, will become zero when the implied vectors are the same:

$$|\mathbf{a}\rangle = |\mathbf{b}\rangle \rightarrow h_{ab} = 1 \rightarrow d_{ab} = 0,$$

and will become more nearby the unit as h_{ab} becomes almost null. The nullity of the Hodgkin and Richards SI is an instance, which will only occur when both compared vectors are orthogonal, a characteristic which will never occur in quantum similarity calculations.

One must be aware that the Carbó SI, being equivalent to a cosine of the angle subtended by the involved vectors, behaves in the contrary way: it becomes unit when both vectors are the same or homothetic and zero in the most dissimilar situation of being orthogonal, a situation that will never occur too when comparing density functions.

Realizing these previous properties, one can see that in order to preserve the geometric structure and behavior of the Hodgkin and Richards SI within a Euclidian distance definition, it is advisable that such a SI be used within this scaled Euclidian distance context, using a formulation as expressed in Eq. (8). Taking it just as the bare ratio between the scalar product and the arithmetic mean of the Euclidian norms of

the involved vectors, although the resultant expression indeed corresponds to a SI, which behaves as the Carbó SI, the fact is that it is out of context use. In this circumstance Hodgkin and Richards SI loses completely its Euclidian distance attachment, hence misses an adequate geometrical significance and has no possible geometrical interpretation whatsoever.

Therefore, summarizing this situation, one can say that Hodgkin and Richards SI, being a basic part of the Euclidean distance expression, has to be used simply not as an SI, but as a DI instead, in order to preserve its geometric origin and signification.

5 Conclusions

Although Hodgkin and Richards SI was originally described as a variation of Carbó SI by these authors, the inclusion of this index in the expression of the Euclidian distance or its role in the scaled Euclidian distance, precludes that it is not suitable to use it out of such DI expressions, but forming part of them.

Euclidian distances do not possess the generalization power and invariance characteristics of Carbó SI, being essentially a binary quantum object comparison tool, but at the same time they do not lack of interest for molecular taxonomy purposes.

Perhaps the new index described adequately here can be called Carbó-Hodgkin-Richards scaled distance DI.

References

1. R. Carbó, L. Leyda, M. Arnau, Int. J. Quantum Chem. **17**, 1185 (1980)
2. R. Carbó, E. Besalú, B. Calabuig, L. Vera, Adv. Quantum Chem. **25**, 253 (1994)
3. R. Carbó, E. Besalú, Theoretical Foundation of Quantum Similarity in *Molecular Similarity and Reactivity: From Quantum Chemical to Phenomenological Approaches,* ed. by R. Carbó. Understanding Chemical Reactivity, vol 14 (Kluwer, Amsterdam, 1995), pp. 3–30
4. R. Carbó, E. Besalú, L. Amat, X. Fradera, J. Math. Chem. **19**, 47 (1996)
5. E.E. Hodgkin, W.G. Richards, Int. J. Quantum Chem. S **14**, 105 (1987)
6. R. Carbó-Dorca, J. Mol. Struct. Teochem **537**, 41 (2001)
7. R. Carbó-Dorca, J. Math. Chem. **32**, 201 (2002)
8. P. Bultinck, R. Carbó-Dorca, J. Math. Chem. **36**, 191 (2004)
9. P. Bultinck, X. Girones, and R. Carbó-Dorca, in *Molecular Quantum Similarity: Theory and Applications*, ed. by K.B. Lipkowitz, R. Larter, T. Cundari. Reviews in Computational Chemistry, vol 21 (Wiley, Hoboken, 2005), pp. 127–207
10. R. Carbó-Dorca, J. Math. Chem. **44**, 628 (2008)
11. R. Carbó-Dorca, J. Math. Chem. **47**, 331 (2010)
12. W. Ayers Paul, R. Carbó-Dorca, J. Math. Chem. **49**, 6 (2011)
13. L.D. Mercado, R. Carbó-Dorca, Quantum similarity and discrete representation of Molecular sets. J. Math. Chem. **49**, 1558–1572 (2011)
14. R. Carbó-Dorca, Quantum similarity, Volume Functions and Generalized Carbó Indices. J. Math. Chem. **49**, 2109–2115 (2011)
15. R. Carbó-Dorca, E. Besalú, Shells, point cloud huts, generalized scalar products, cosines and similarity tensor representations in vector semispaces. J. Math. Chem. (2011). doi:10.1007/s10910-011-9906-4
16. P. Bultinck, R. Carbó-Dorca, J. Chem. Inf. Comp. Sci. **43**, 170 (2003)
17. P.H.A. Sneath, R.R. Sokal, *Numerical Taxonomy* (W. H. Freeman & Co., San Francisco, 1973)
18. M.M. Deza, E. Deza, *Encyclopedia of Distances* (Springer, Berlin, 2009)