

## Molecular quantum similarity using conceptual DFT descriptors

PATRICK BULTINCK<sup>1,\*</sup> and RAMON CARBÓ-DORCA<sup>1,2</sup>

<sup>1</sup>Department of Inorganic and Physical Chemistry, Ghent University, Krijgslaan 281, B-9000 Gent, Belgium

<sup>2</sup>Institute of Computational Chemistry, University of Girona, Campus de Montilivi, 17005 Girona, Spain  
e-mail: Patrick.Bultinck@UGent.be; quantumqsar@hotmail.com

**Abstract.** This paper reports a Molecular Quantum Similarity study for a set of congeneric steroid molecules, using as basic similarity descriptors electron density  $\rho(\mathbf{r})$ , shape function  $S(\mathbf{r})$ , the Fukui functions  $f^+(\mathbf{r})$  and  $f^-(\mathbf{r})$  and local softness  $s^+(\mathbf{r})$  and  $s^-(\mathbf{r})$ . Correlations are investigated between similarity indices for each couple of descriptors used and compared to assess whether these different descriptors sample different information and to investigate what information is revealed by each descriptor.

**Keywords.** Molecular quantum similarity; steroid; Fukui function; shape function.

### 1. Introduction

An essential feature of human observations is the fact that they rely on comparison and classification to interpret observations. According to Rouvray,<sup>1</sup> all issues of comparison, and thus of classification, are related to similarity, stressing the ubiquitous nature of the similarity concept. A classical example of the ubiquitous importance of similarity in interpreting observations is that of human facial expressions, where one interprets some expressions as indicating happiness and others sadness. Chemistry, as another human activity, is no exception to this ubiquitous nature of similarity. Concepts like acidity and basicity rely on classification; the periodic system is a clear example of grouping atoms together based on similarity etc.

The above statements clearly justify developing schemes to examine the similarity between molecules so that they can be classified. The interest in developing such schemes goes back to the earliest days of chemistry, and continues to be growing. One result of this long history is that there have been developed a very large number of ways to assess the similarity of molecules. Another reason for this wealth of similarity approaches is the fact that, as chemical knowledge grew; new concepts have entered the field. A third reason which should not be underestimated is the fact that according to Herndon and Bertz:<sup>2</sup> “Similarity, like beauty, lies in the eyes of the beholder”. This means for example that an organic

chemist may use other concepts to classify molecules than a quantum chemist, and that a physical chemist might use even other concepts. So the extent of molecular similarity will depend on the concept used. A molecular property that acts as such a concept and that describes a molecule is called a molecular descriptor. Over time many different molecular descriptors have been developed, especially in e.g. computational medicinal chemistry where hundreds of such descriptors are used. Following Downs,<sup>3</sup> these descriptors can be grouped in different categories, depending on their dimensionality. As such, one can distinguish feature counts as a first category. There the molecular descriptors are simply counts of specific features of a molecule, e.g. the number of hydrogen bond acceptor atoms. In another level of complexity one may use physicochemical parameters as descriptors, e.g. the well-known Log P. Topological and topographical indices are also a well-known class of indices, including the Wiener index, the Balaban index, the indices introduced by Randić, the Zagreb index and the Hosoya index.<sup>4–13</sup> The ultimate group of descriptors consists of so-called field descriptors. Examples of fields are the electron density, steric fields, electrostatic potentials, hydrophobic fields and so on.

This paper aims at using different field based descriptors to assess what different information is obtained from them. More precisely we wish to examine the use of the total electron density, shape function, Fukui functions and other fields functions rooted in conceptual DFT.<sup>14–16</sup> Several of such descriptors have been used individually in other studies, but a

\*For correspondence

comparison using dendrograms for a well-established and related set of molecules has yet to appear. It is well worth mentioning in this context the work of Boon *et al*<sup>17-20</sup> who addressed already several of these field descriptors for sets of peptide isosteres. No in-depth study seems to have been performed yet on the resulting changes in similarity ordering in a larger set of chemically related molecules. Inspired also on the recent statement of De Proft *et al*<sup>21</sup> that similarity in shape is a fundamental issue to be looked at, we set out to make a study of similarity among a set of molecules using different field descriptors. This study will focus on a set of congeneric molecules, since there one finds the most critical applications of quantum similarity.

## 2. Molecular quantum similarity and field descriptors

### 2.1 Molecular quantum similarity

The theory of molecular quantum similarity (MQS) has been reviewed in detail in several papers, so the reader is referred to references.<sup>22-29</sup> Most recently MQS and its different fields of application, including some other field descriptors, have been reviewed by Bultinck *et al*.<sup>30</sup>

Molecular quantum similarity is concerned with the quantification of similarity between two molecules A and B via the evaluation of so-called Molecular quantum similarity measures (MQSM):

$$Z_{AB} = \int F_A(\mathbf{r}_1)\Omega(\mathbf{r}_1, \mathbf{r}_2)F_B(\mathbf{r}_2)d\mathbf{r}_1d\mathbf{r}_2. \quad (1)$$

$F_A$  is the field descriptor used for molecule A, which in most applications up to the present, corresponds to the electron density of A.  $\Omega(\mathbf{r}_1, \mathbf{r}_2)$  can be any positive definite operator, such as the Coulomb operator  $(\mathbf{r}_1 - \mathbf{r}_2)^{-1}$ , the gravitational operator  $(\mathbf{r}_1 - \mathbf{r}_2)^{-2}$  and the most often used operator:  $\delta(\mathbf{r}_1 - \mathbf{r}_2)$ . The latter is the Dirac delta function, which turns the MQSM of (1) into an overlap measure between the two functions involved, that is:

$$Z_{AB} = \int F_A(\mathbf{r}_1)F_B(\mathbf{r}_1)d\mathbf{r}_1. \quad (2)$$

It is these MQSM that will be used throughout the present paper. The quantum similarity between molecules A and B can then easily be quantified through a Euclidean distance between the infinite dimen-

sional density vectors or related conceptual DFT vectors, that is:

$$d_{AB}^2 = Z_{AA} + Z_{BB} - 2Z_{AB}. \quad (3)$$

The  $Z_{AA}$  and  $Z_{BB}$  are in this context the overlap integrals over the field descriptor for twice the same molecule, and are called self-similarity measures. These have been found to correlate well with several physicochemical molecular properties<sup>31-35</sup> and have been found to be a measure of the electronic charge density concentration in molecules.<sup>36</sup> Another similarity index that has found wide application is the so-called Carbó index  $C_{AB}$ <sup>37</sup> which is a generalized cosine, giving a value between 0 and 1, the latter indicating perfect similarity.

$$C_{AB} = Z_{AB}/(Z_{AA}Z_{BB})^{1/2}. \quad (4)$$

Once the similarity measures have been obtained, a molecular quantum similarity matrix for a set of  $N$  molecules can be constructed as:

$$\mathbf{Z} = \begin{bmatrix} Z_{11} & \dots & Z_{1N} \\ \vdots & \ddots & \vdots \\ Z_{N1} & \dots & Z_{NN} \end{bmatrix}. \quad (5)$$

Each of the columns or the rows of this matrix can be considered a discrete representation the corresponding molecular field descriptor in the subspace formed by all fields of the  $N$  molecules. Then, the  $i$ 'th column of  $\mathbf{Z}$ , denoted  $z_i$ , can be considered as a discrete representation of molecule  $i$  in the space spanned by the field descriptors of the  $N$  molecules.<sup>29</sup> In this sense these column vectors discretize the infinite dimensional representations of the molecular fields into an  $N$  dimensional vector representation, where all the numbers in the vector are real, positive definite values. The columns of the molecular quantum similarity matrices can be associated in turn to molecular descriptors. These molecular vector representations have two important special properties. The vector descriptors are universal in the sense that they can be obtained for any molecule of the set and from any molecular set. Furthermore they are unbiased except in the stage of the selection of the operator used to evaluate the MQSM in (1). In the remainder of the paper, the similarity matrices (5) obtained using different field descriptors will be compared. From the matrices with the different field

descriptors it is also possible to draw sequential agglomerative hierarchical non-overlapping (SAHN) dendrograms as described by Bultinck *et al.*<sup>38</sup> Briefly, these SAHN dendrograms first identify the most similar pair of molecules in the similarity matrix. These are clustered together, and a new averaged density function is constructed. The similarity matrix is then reconstructed, meaning that the similarity is computed for all molecules versus the averaged density function. Then, the new most similar pair is clustered, including the possibility of the most similar pair being a combination with a newly constructed averaged density function. Such dendrograms allow a graphical inspection of the degree of similarity between molecules and allow grouping molecules together in different sequential steps.

It is well known that the MQSM are dependent on the molecular alignment as is immediately clear from (1). This dependence may cause important problems when comparing the similarity between different pairs of molecules. Although algorithms have been proposed to solve the molecular superposition problem, most quantum molecular similarity applications still require molecular alignment to be performed. This may be done in different ways; one of the soundest ones can be performed via maximizing MQSM, as is done in the MaxiSim<sup>39</sup> and QSSA<sup>40</sup> algorithms. However, when using the total electron density this alignment is dominated by the tendency to superpose atomic nuclei. In principle, one should then for every field descriptor maximize the MQSM between every two pairs of molecules. This in itself would be very interesting since it might reveal other alignments. Unfortunately, maximizing the MQSM is not computationally straightforward in general for any field descriptor. Therefore, a structural alignment procedure is used. Then after knowing how molecular pairs are aligned, the MQSM are computed to assess the different information contained in the similarity matrix for different field descriptors.

## 2.2 Field descriptors from conceptual DFT

So far, not very deep insight has been put forward about the use of diverse field descriptors in molecular quantum similarity. Until now the single most often-used field descriptor has been the electron density  $\mathbf{r}_A(\mathbf{r})$ . Application of this descriptor in molecular quantum similarity has been found to give important new insights in many different applications, including quantitative structure-activity relationships

(QSAR),<sup>29,41</sup> chirality<sup>19,20,42</sup> and many more diverse fields.

In the actual calculations performed in this work, use will be made of DFT calculations, where the electron densities needed are given by the classical expression:

$$\mathbf{r}_A = \sum_{nm} D_{nm} \mathbf{j}_n \mathbf{j}_m^* \quad (6)$$

where  $D$  is the charge and bond order matrix and the  $\{\mathbf{j}_v\}$  are the basis functions used in the molecular SCF procedure. The MQSM then are given by:

$$Z_{AB} = \sum_{n \in A} \sum_{m \in A} \sum_{d \in B} \sum_{l \in B} D_{nm} D_{dl} \int \mathbf{j}_m^*(\mathbf{r}) \mathbf{j}_n(\mathbf{r}) \mathbf{j}_l^*(\mathbf{r}) \mathbf{j}_d(\mathbf{r}) d\mathbf{r}. \quad (7)$$

Such four basis function overlap integrals as described in (7) can be calculated using the classical approaches for the evaluation of overlap integrals over Gaussian type orbitals.

The electron density, however, is not the only imaginable field descriptor. Another related descriptor is the so-called shape function  $\mathbf{s}_A(\mathbf{r})$ .<sup>43</sup> This is obtained simply as:

$$\mathbf{s}_A(\mathbf{r}) = N_A^{-1} \mathbf{r}_A(\mathbf{r}), \quad (8)$$

where  $N_A$  is the number of electrons in the molecule A. The shape function, as the density itself, determines every observable for the system, and somewhat unexpectedly can be shown to hold information on the total number of electrons despite that for all molecules:

$$\int \mathbf{s}_A(\mathbf{r}) d\mathbf{r} = 1. \quad (9)$$

Moreover, it has been shown that the shape function is related to several DFT based reactivity indices and a variational procedure can be derived to obtain the energy of the system.<sup>44-45</sup> Bultinck *et al* have shown that the shape function, just as the density function, belongs to vector semispaces unit shell, and that given the simple relationship (8), they should hold very similar information.<sup>45</sup> Such reasoning is however based on single molecule considerations. Inspired on the suggestion of De Proft *et al.*,<sup>21</sup> the similarity ordering of the Cramer steroid set

based on shape functions molecular quantum similarity measures will be compared to that obtained with the density function.

Substituting (8) in the Carbó similarity index of (4) immediately permits to show that this index remains invariant. Although other similarity measures do change like the Euclidean distance in (3), as does e.g. the Hodgkin–Richards index  $H_{AB}$ .<sup>46,47</sup>

$$H_{AB} = 2Z_{AB}/(Z_{AA} + Z_{BB}). \quad (10)$$

The noninvariance of the Hodgkin–Richards index is not surprising as it is *not* attached, as the Carbó index is, to any geometrical feature of the molecular cloud.

Due to these considerations, in this work we have chosen to compare the different similarity indices with both the density function and shape function as field descriptors. The Euclidean distance as defined in (3) has been also used for comparative purposes. The MQSM expression over shape functions is naturally very similar to the one in (7).

As was mentioned above, the MQSM are often dominated by the inner core density of the atoms. It is, however, well known that chemical phenomena are mostly related to small differences in density in the valence region. A very often used concept used for reactivity studies based mainly on the valence density, is that of the Fukui function,<sup>14–16</sup> introduced originally by Parr and Yang.<sup>48</sup> In conceptual DFT and applying a finite difference approximation, this is expressed as:

$$f_A^+(\mathbf{r}) = \mathbf{r}_A^{N_A+1}(\mathbf{r}) - \mathbf{r}_A^{N_A}(\mathbf{r}), \quad (11)$$

$$f_A^-(\mathbf{r}) = \mathbf{r}_A^{N_A}(\mathbf{r}) - \mathbf{r}_A^{N_A-1}(\mathbf{r}), \quad (12)$$

$$f_A^0(\mathbf{r}) = \frac{1}{2}(f_A^+(\mathbf{r}) + f_A^-(\mathbf{r})), \quad (13)$$

where the superscripts denote the density of the molecule with  $N_A$  electrons,  $N_A - 1$  and  $N_A + 1$  electrons. Equations (11)–(12) are actually exact in exact DFT.<sup>49,50</sup> The densities are all obtained via quantum chemical calculation, all on the same geometry since the Fukui function is defined under constant external potential:<sup>14–16</sup>

$$f_A(\mathbf{r}) = (\partial \mathbf{r}_A(\mathbf{r}) / \partial N)_{\text{vext}}. \quad (14)$$

It is straightforward to program the necessary Fukui function similarity indices for all Fukui functions.

Fukui functions are important reactivity descriptors because they indicate the preferred locations for electrophilic and nucleophilic reactions. Moreover, given their defining equations one has as requirement that:

$$\int f_A(\mathbf{r}) d\mathbf{r} = 1. \quad (15)$$

The use of the Fukui function similarity between different molecules is, however, a dubious question,<sup>51</sup> with the Fukui function being only useful for intramolecular comparison. Nevertheless, it will be included in the present study because, as will be shown below, the similarity in local softness will also include Fukui functions. It is a simple matter to obtain Fukui function similarity with local softness similarity calculations.

Another often used concept that is interesting to explore is the local softness.<sup>14–16</sup> This reactivity descriptor, introduced by Yang and Parr,<sup>52</sup> has also been shown to give interesting new information, and its application in molecular quantum similarity has not yet been explored in great detail. The local softness can be obtained as:

$$s_A^+(\mathbf{r}) = f_A^+(\mathbf{r}) \cdot S, \quad (16)$$

$$s_A^-(\mathbf{r}) = f_A^-(\mathbf{r}) \cdot S, \quad (17)$$

$$s_A^0(\mathbf{r}) = f_A^0(\mathbf{r}) \cdot S, \quad (18)$$

where  $S$  is a global property, namely the global softness, which can be approximated as:

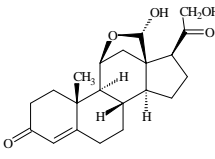
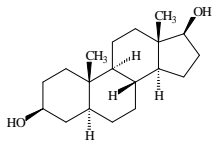
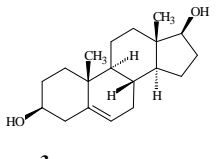
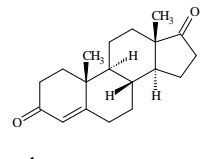
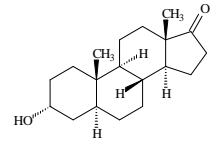
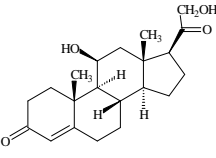
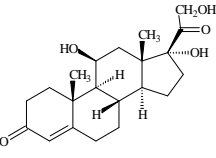
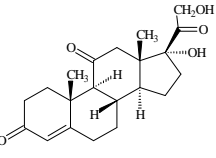
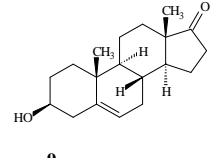
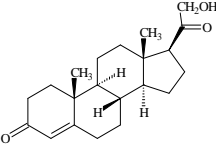
$$S = 1/(\text{IE} - \text{EA}) \quad (19)$$

where IE and EA are respectively the molecular ionization energy and the electron affinity. Given the value of  $S$ , it is again fairly simple to compute the necessary MQSM. Both Koopmans theorem and separate calculations for the ionic species have been used for calculating the global softness.

### 3. Computational methods

For the analysis of the differences in computed similarity matrices with the different field descriptors shown above, it was used a subset of the globulin binding steroids first described by Cramer *et al*<sup>53</sup> also employed to develop QSAR models<sup>54</sup> afterwards. The complete dataset of Cramer *et al.* has also been previously used in molecular quantum similarity studies and to develop quantum QSAR models.<sup>55,56</sup> The 10

**Table 1.** The set of steroids considered in the present example application.

			
1 Aldosterone	2 Androstanediol	3 5-Androstanediol	4 4-Androstenedione
			
5 Androsterone	6 Corticosterone	7 Cortisol	8 Cortisone
			
9 Dehydroepiandrosterone	10 11-Deoxycorticosterone		

molecules included in the present set are shown in table 1.

The 3-D structures of all molecules were generated using AM1<sup>57,58</sup> geometry optimizations. Electron densities were then obtained using B3LYP<sup>59–61</sup>/6-31G\* single-point calculations for all species involved. Once the electron density is known, the MQSM for the different field descriptors were calculated using an in-house written program that extracts all necessary information from the Gaussian03<sup>62</sup> DFT calculations. Four center overlap integrals were computed using the classical methods.<sup>63</sup>

As was mentioned previously, molecular alignment plays an important role in determining the values of the MQSM. In the present study, molecular alignment was performed using the TGSA structural alignment algorithm.<sup>64</sup> Once the molecular similarity indices were calculated, dendrograms were constructed to reveal molecular relationships in the way described previously by Bultinck *et al.*<sup>38</sup>

#### 4. Results and discussion

Prior to the discussion of the results obtained, it is worth generalizing how to obtain the necessary similar-

ity indices from a single similarity calculation and to discuss the importance and consequences of the homothecies existing between many conceptual DFT quantities. The MQSM are usually evaluated with the electron density as the most important descriptor. That is:

$$Z_{AB}^r = \int \mathbf{r}_A(\mathbf{r}_1) \mathbf{r}_B(\mathbf{r}_1) d\mathbf{r}_1. \quad (20)$$

The Euclidean distance using the electron density is then obtained as:

$$d_{AB}^{2,r} = Z_{AA}^r + Z_{BB}^r - 2Z_{AB}^r. \quad (21)$$

It would be beneficial if one could compute the necessary MQSM for different descriptors at once. For instance, it is trivial to show that the shape function based similarity matrix  $\mathbf{Z}^s$  is related to the density function similarity matrix  $\mathbf{Z}^r$  by the matrix equation:

$$\mathbf{Z}^r = \mathbf{N} \mathbf{Z}^s \mathbf{N} \quad (22)$$

where  $\mathbf{N}$  is a diagonal matrix with elements  $N_i$  equal to the number of electrons in molecule  $i$ . Similarly, the same goes for the interrelation between the Fu-

kui similarity matrix and the softness similarity matrix but using the diagonal matrix with the global softness as diagonal elements. The Euclidean distances are then dependent on  $\mathbf{N}$ , since one has, for example, for the shape function:

$$d_{AB}^{2,r} = (\mathbf{N}\mathbf{Z}^s\mathbf{N})_{AA} + (\mathbf{N}\mathbf{Z}^s\mathbf{N})_{BB} - 2(\mathbf{N}\mathbf{Z}^s\mathbf{N})_{AB}. \quad (23)$$

So the nature of  $\mathbf{N}$  will influence the Euclidean distances. However, as was discussed in more detail by Bultinck *et al*<sup>45</sup> the homothety relationship between the density and shape function, leaves invariant the Carbó index. Two vectors are homothetic whenever they are related by a simple scaling. So one has the following relationships for the Carbó indices  $C_{AB}^f$  with different descriptors:

$$C_{AB}^r = C_{AB}^s \quad \text{and} \quad C_{AB}^f = C_{AB}^s. \quad (24)$$

In order to see the effect of introducing the dependence on  $\mathbf{N}$  on the similarity between molecules, it has been chosen the Euclidean distance as a D-class measure rather than a C-class descriptor. The latter namely come in many different forms, some that give the same similarity value when using homothetic descriptors, others that do not. The D-class measure, however, is a very simple way to introduce the dependence on  $\mathbf{N}$ , and allows examining its influence on similarity ordering. Naturally, there also exist relationships between the different Euclidean distances. As an example, one can deduce that the relationship between density function and shape function-based distances is:

$$d_{AB}^{2,s} = (N_A N_B)^{-1} d_{AB}^{2,r} - \left[ \frac{N_A - N_B}{N_B} Z_{AA}^s + \frac{N_B - N_A}{N_A} Z_{BB}^s \right]. \quad (25)$$

This shows that the Euclidean distance for density functions is related to that in shape functions by a scaling relation and addition of weighed self similarities. It is then also clear that the relative similarity ordering may be changed upon going from density functions to shape functions.

For the calculation of the Fukui index similarity, and via a relationship as (22) also for the local softness similarity, a finite difference approach is used through the DFT calculation for the neutral molecule, and the singly charged ions.

Coming back to electron density based MQSM; the following observations can be made. First of all, it should be noted that there is no substantial difference in the similarity ordering using the different similarity indices. This is illustrated in figure 1.

This is easily explained since there are simple relationships between virtually all similarity indices in common use.<sup>65-68</sup> For example between the Carbó and Hodgkin–Richards indices one can write:

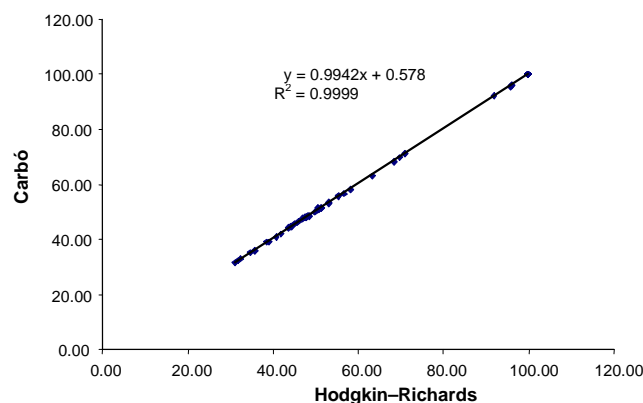
$$H_{ij} = 2\mathbf{a}^{1/2}/(1 + \mathbf{a})C_{ij}, \quad (26)$$

where, following Maggiora *et al*<sup>65</sup>  $\mathbf{a}$  is specific for each combination A, B:

$$\mathbf{a} = \min(Z_{AA}, Z_{BB})/\max(Z_{AA}, Z_{BB}). \quad (27)$$

This explains why in relatively congener molecules, one should not expect  $\mathbf{a}$  to deviate very far from 1. The lowest value in the present study occurs between molecules 1 and 3, and  $\mathbf{a} = 0.71$ . The factor  $2\sqrt{\mathbf{a}}/(1 + \mathbf{a})$  behaves in such a way that a value of 0.71 results in a coefficient of 0.98, so not much difference is to be expected between both similarity indices. This is also clear from figure 2.

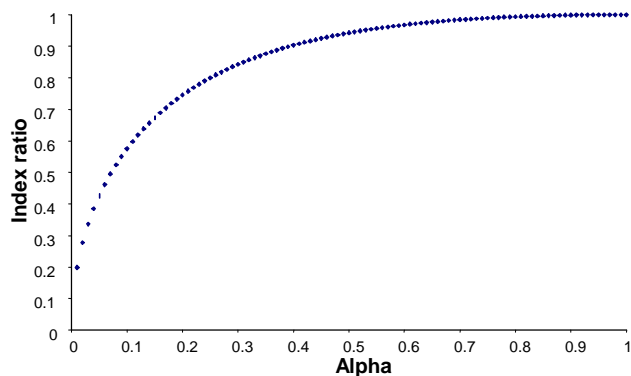
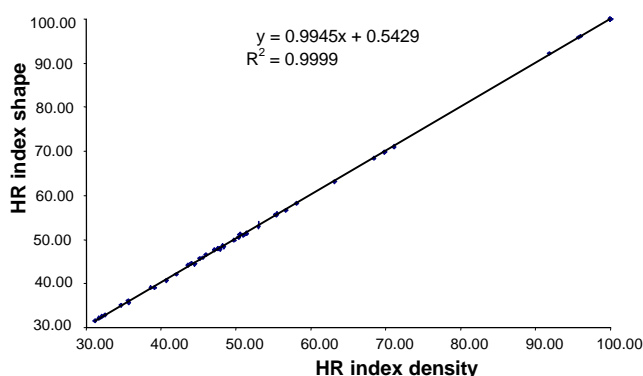
This shows clearly that the difference between the Hodgkin–Richards and Carbó indices will become relevant only for molecular sets containing quite different molecules, meaning that self similarities can have wide-spread values. In most applications, however, the value of  $\mathbf{a}$  is always quite high (above 0.5, meaning a ratio of above 0.94), so that figure 2 clearly shows that the similarity indices will not differ very much. The self-similarities of the molecules in the present set are shown in table 2.



**Figure 1.** Carbó index versus Hodgkin–Richards index for electron density-based similarity for the set of steroids.

**Table 2.** Self-similarities for all steroids, based on electron density.

Molecule number	Self-similarity	Molecule number	Self-similarity
1	1063.57	6	983.07
2	758.58	7	1063.76
3	758.55	8	1063.64
4	758.59	9	758.59
5	758.62	10	902.20

**Figure 2.** Plot of the index ratio  $2a^{1/2}/(1+a)$  versus  $a$ .**Figure 3.** Correlation between Hodgkin-Richards similarity index computed with the electron density and shape function.

The presence of several molecules with very near self-similarities is typical of medicinal chemistry environments, where often molecules with related skeletons are encountered.

Naturally, there are also correlations between the C-class indices and the Euclidean distance measure. Such relationship is less analytical as between different C-class indices, but nevertheless, one finds a correlation coefficient of 97% between the Carbó index and the Euclidean distance. This is not surpris-

ing as distances and Carbó indices within a molecular set can be transformed one into another.

It is sometimes argued that the use of the shape function offers important advantages over the use of the electron density. The first question to be settled in quantum similarity is thus to investigate whether shape function and electron density MQSM do effectively show different information content. The data set used in the present study contains molecules with between 156 and 196 electrons. As was commented above, no change in the Carbó-index occurs when going from electron density to the shape function. When looking for example at the Hodgkin-Richards index, somewhat unexpectedly no substantial differences in the quantum similarity are observed. This is illustrated in figure 3.

As seen, there is nearly perfect correlation with a unit slope. There are clearly no outliers, despite a difference of 40 electrons between two molecules. Taking as an example molecules 4 (156 electrons) and 7 (196 electrons), the similarity index using the electron density and shape similarity differs less than 1%. Denoting the electron density based Hodgkin-Richards index as  $H_{AB}^r$  and that based on the shape function as  $H_{AB}^s$ , their ratio is given by:

$$\frac{H_{AB}^r}{H_{AB}^s} = \frac{x^{-1}Z_{AA}^r + xZ_{BB}^r}{Z_{AA}^r + Z_{BB}^r}, \quad (28)$$

where  $x$  is the ratio of the numbers of electrons of A and B:

$$x = N_A/N_B. \quad (29)$$

From the numerical data, it is found that  $H_{AB}^r/H_{AB}^s$  is usually quite close to unity. The deeper reason for this is naturally that the self-similarity scales directly with the number of electrons. So, it can be concluded, on the basis of the analysis of the C-class indices behavior, that shape functions do not provide new information when compared with the homothetic

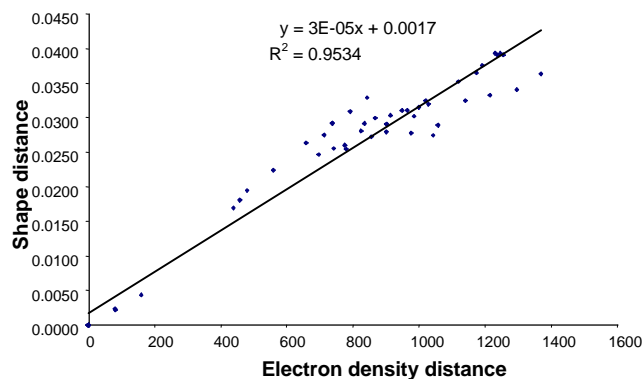
electron densities. Also one can say that there is no real influence of the rescaling of the density function to shape function for both C-class indices. The D-class index does however show bigger changes. This is depicted in figure 4.

The reason for the less pronounced correlation appears as a natural consequence of the fact that the relationship between distances with shape and density functions is not so straightforward, see (25). The ratio of both distances is given by:

$$\frac{d_{AB}^r}{d_{AB}^s} = \frac{N_A N_B (Z_{AA} + Z_{BB} - 2Z_{AB})}{(x^{-1}Z_{AA} + xZ_{BB} - 2Z_{AB})}, \quad (30)$$

with  $x$  again given by (29). From a statistical point of view, the correlation coefficient between the shape function and electron density based Euclidean distances is such that one cannot conclude the shape function and electron densities offer new or different information in case of quantum similarity within a set of congener molecules.

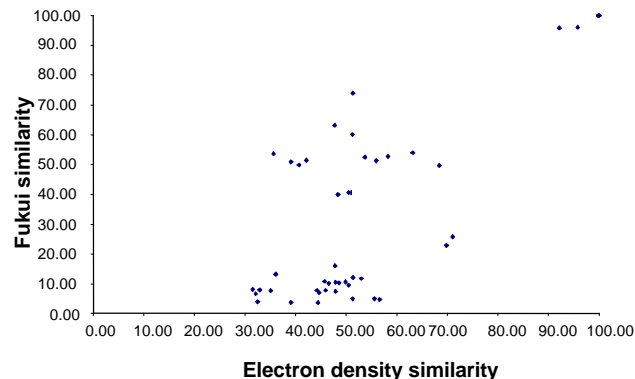
It is interesting to examine whether Fukui functions do indeed offer other information than the electron density. The Fukui densities defined through (11)–(13) could be expected to offer other, additional information on the similarity between molecules. In the electron density and the shape function, there is a dominance of the core regions. In the Fukui functions, the core region densities are largely cancelled since the ionization of the molecules does not largely influence the core regions. It is thus plausible that they will provide insight largely on valence effects. Figure 5 shows a plot of the Carbó indices based on the  $f_A^-(\mathbf{r})$  Fukui function versus those based on the electron density.



**Figure 4.** Euclidean distances between steroid molecules using electron density functions and shape functions.

This figure clearly shows that the Fukui functions certainly offer other information than density functions. The correlation is very low, namely 72%, meaning loss of any correlation. Within the set of congener molecules, it is worth examining the relationship between the  $f_A^-(\mathbf{r})$  and  $f_A^+(\mathbf{r})$  index. The correlation index between the Carbó index based on both indices is 93%. This means that both indices sample slightly different similarity information. Again the most discriminating similarity index is the Euclidean distance.

It has been suggested that the local softness would be the most advisable property for intermolecular reactivity comparison. As mentioned above, the Carbó index does not change when going from the Fukui function to local softness, since the latter is simply a scaled Fukui function. We have investigated the correlation between the  $f_A^+(\mathbf{r})$  based Euclidean distance and the  $S_A^+(\mathbf{r})$  based Euclidean distance. The correlation coefficient in this case is approximately 95%. At first glance, this would seem to infer that the local softness and Fukui functions do not yield different information for the study of molecular similarity. The reason is that a similar relationship exists between both indices as in (30). This is a fortiori true for the Hodgkin–Richards similarity indices, where similar relationships exist as in (28), using now the ratio of the global softness values  $x = S_A/S_B$ . Another observation is that within the set considered here, the use of the Koopmans theorem or the effectively calculated ionization energies and electron affinities, does cause some differences. It was found that for the correlation mentioned above, the correlation coefficients can drop to 85%, allowing us to conclude that some different information is contained



**Figure 5.** Correlation between Carbó similarity indices based on electron density and Fukui index  $f_A^-(\mathbf{r})$ .



in both indices, at least when using Euclidean distances. All this supports the previous conclusion of Geerlings and co-workers.<sup>51</sup> Suppose that one compares the reactivity of a functional group between two very different molecules; the Fukui function on this reactive region in one molecule may be smaller than in the other molecule. Through the scaling with the global softness, one can find that the local softness is bigger in the first molecule, indicating higher reactivity. It is clear that for such subtle comparisons, the Euclidean distances are the most sensitive similarity indices. For the C-class descriptors, it would be preferable to use the Carbó index, since this has a clear geometrical background. The fact that Carbó indices do not differ, can be seen as a manifestation of the fact that the shape function and the density function contain the same information. This in turn then also explains why a variational procedure can be constructed from both functions.<sup>21,44,45</sup> The interpretation of the Euclidean distance measure is slightly more difficult since these indices have unbound upper limits, whereas C-class indices are contained in the interval ]0,1].

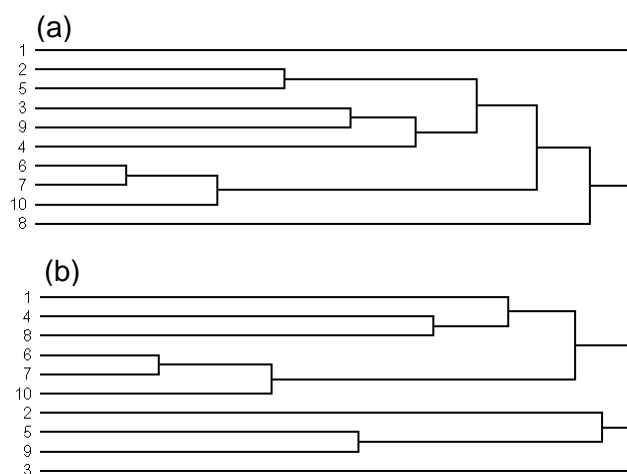
As a general conclusion, one notices that the Carbó index, as well as the Hodgkin–Richards index is insensitive to the use of the electron density or the shape function. A similar conclusion may be drawn for the Fukui function versus local softness. Even when the number of electrons and the global softness in the molecules can differ substantially, the ratio between the Euclidean distances is too small to drastically change the similarity orderings. Yet, the latter similarity indices show the largest differences. It is then up to the researcher to decide whether to use a function holding inherently more (direct) information about the molecule or to use a scaled quantity that does not (directly) contain specific molecular information. For the issue for molecular quantum similarity, this does not bring about large changes.

After having considered the correlations between different descriptors, we now turn to the discussion of the similarity ordering between the molecules. Such a study is most easily carried out using similarity dendrograms, such as the quantum similarity agglomerative clustering hierarchical networks of Bultinck *et al.*<sup>38</sup> Given the very good correspondence between Euclidean distances based on the use of electron density versus that based on the shape function, one observes virtually no difference in the dendrograms. For a thorough discussion of the dendrograms for the steroid set, the reader is referred to

Bultinck *et al.*<sup>38</sup> A more interesting comparison lies in the study of similarity indices based on the density function and local softness, in this case  $S_A^-(\mathbf{r})$ . Figure 6 shows both resulting dendrograms.

It is immediately seen that the clustering does differ to an important extent. Different clusters are clearly formed. In both dendrograms the structurally most similar molecules are gathered first. Interesting cases appear, however, like for the molecular couples 4–8 and 5–9. In the softness dendrogram, molecule 9 is the closest molecule to 5, whereas this is not the case for the density based dendrogram. Also notice the special behavior of molecule 3. All this makes clear that the softness, here illustrated via  $S_A^-(\mathbf{r})$ , does indeed sample completely different information.

It remains, however, a difficult task to conclude which of the descriptors has the largest information content. To arrive to such a conclusion, one needs to have neutral reference data, such as biological activities for the steroid molecules such as those considered here. Quantum QSAR models based on the electron density similarity measures have already been found to yield good QSAR models. The development and assessment of QSAR models based on similarity matrices using the local softness are, however, not yet available. From the above findings, one can certainly envision that the latter similarity matrices could act as a new set of independent molecular descriptors, thereby extending the range of discretized molecular descriptors which are derived within quantum QSAR.



**Figure 6.** Dendrograms obtained from the Carbó index based on the density function (a) and on the local softness (b).

## 5. Conclusion

By performing molecular quantum similarity studies, it has been found that different field descriptors from conceptual DFT contain different chemical information, but also that some of these descriptors are redundant. The shape function and the density function largely give the same quantum similarity information. This agrees with previous findings of Ayers *et al* that both possess quite related information, although in a less straightforward manner for the shape function. The Carbó index very clearly reveals this point, although it was shown that the Hodgkin–Richards index does not change much either when going from the density function to shape function. This is due to the existence of simple relations between the C-class correlation indices when using different descriptors where one descriptor is simply a rescaling of another.

The present study has shown that, within the boundaries of a set of congeneric molecules, the density function and local softness can be used as the most descriptive entities, with the shape function and the Fukui function giving very similar information. All this also agrees with the homothety relationships described by Bultinck *et al*.

## Acknowledgements

PB acknowledges the support for the Fund for Scientific Research, Flanders for continuous support to the quantum chemistry research group. RCD expresses his acknowledgement to the Ministerio de Ciencia y Tecnología for a grant that partially sponsored this work and also for a Salvador de Madariaga fellowship permitting his stay at Gent University.

## References

1. Rouvray D H 1995 *Top. Curr. Chem.* **173** 1
2. Herndon W C and Bertz S H 1987 *J. Comput. Chem.* **8** 367
3. Downs G M 2003 Molecular descriptors. *Computational medicinal chemistry for drug discovery* (eds) P Bultinck, H De Winter, W Langenaeker and J P Tollenaere (New York: Dekker) pp. 364–386
4. Wiener H 1947 *J. Am. Chem. Soc.* **69** 17
5. Balaban A T (ed.) 1976 *Chemical applications of graph theory* (London: Academic Press)
6. Randić M and Wilkins C L 1979 *J. Chem. Inf. Comput. Sci.* **19** 31
7. Ruecker G and Ruecker C 1993 *J. Chem. Inf. Comput. Sci.* **33** 683
8. Randić M 1990 Design of molecules with desired properties. *Concepts and applications of molecular similarity* (eds) M A Johnson and G M Maggiora (New York: Wiley-Interscience) pp. 77–145
9. Randić M 1975 *J. Am. Chem. Soc.* **97** 6609
10. Kier L B and Hall L H 1986 *Molecular connectivity in structure-activity analysis* (New York: John Wiley and Sons)
11. Gutman I and Trinajstić N 1972 *Chem. Phys. Lett.* **17** 535
12. Nikolić S, Kovacević G, Milicević A and Trinajstić N 2003 *Croat. Chem. Acta* **76** 113
13. Hosoya H 1971 *Bull. Chem. Soc. Japan* **44** 2332
14. Parr R G and Yang W 1989 *Density functional theory of atoms and molecules* (Oxford: University Press)
15. Geerlings P, De Proft F and Langenaeker W 2003 *Chem. Rev.* **103** 1793
16. Chattaraj P K, Nath S and Maiti B 2004 Reactivity descriptors. *Computational medicinal chemistry for drug discovery* (eds) P Bultinck, H De Winter, W Langenaeker and J P Tollenaere (New York: Dekker) pp. 295–322
17. Boon G, De Proft F, Langenaeker W and Geerlings P 1998 *Chem. Phys. Lett.* **295** 122
18. Boon G, Langenaeker W, De Proft F, De Winter H, Tollenaere J P and Geerlings P 2001 *J. Phys. Chem.* **105** 8805
19. Boon G, Van Alsenoy C, De Proft F, Bultinck P and Geerlings P 2003 *J. Phys. Chem.* **A107** 11120
20. Boon G, Van Alsenoy C, De Proft F, Bultinck P and Geerlings P 2005 *J. Mol. Struct. (Theochem.)* **727** 49
21. De Proft F, Ayers P W, Sen K D and Geerlings P 2004 *J. Chem. Phys.* **120** 9969
22. Carbó R and Calabuig B 1990 Molecular similarity and quantum chemistry. *Concepts and applications of molecular similarity* (eds) M A Johnson and G M Maggiora (New York: Wiley-Interscience) pp. 147–171
23. Besalú E, Carbó R, Mestres J and Solà M 1995 *Top. Curr. Chem.* **173** 31
24. Carbó-Dorca R and Besalú E 1998 *J. Mol. Struct. (Theochem.)* **451** 11
25. Carbó-Dorca R, Amat L I, Besalú E and Lobato M 1998 Quantum similarity. *Advances in molecular similarity* (eds) R Carbó-Dorca and P G Mezey (London: JAI Press) vol 2, pp. 1–42
26. Carbó-Dorca R, Amat L I, Besalú E, Gironés X and Robert D 1999 *J. Mol. Struct. (Theochem.)* **504** 181
27. Carbó-Dorca R, Robert D, Amat L I, Gironés X and Besalú E 2000 *Lecture Notes in Chemistry* **73** 1
28. Besalú E, Gironés X, Amat L I and Carbó-Dorca R 2002 *Acc. Chem. Res.* **35** 289
29. Carbó-Dorca R and Gironés X 2004 Quantum similarity and quantitative structure-activity relationships. *Computational medicinal chemistry for drug discovery* (eds) P Bultinck, H De Winter, W Langenaeker and J P Tollenaere (New York: Dekker) pp. 364–386
30. Bultinck P, Gironés X and Carbó-Dorca R 2005 Molecular Quantum similarity: Theory and applications. *Reviews in Computational chemistry* (eds) K Lipkowitz, R Larter and T R Cundari (New York: John Wiley & Sons) vol 21, pp 127–207

31. Amat LI, Carbó-Dorca R and Ponec R 1998 *J. Comput. Chem.* **14** 1575
32. Gironés X, Amat LI and Carbó-Dorca R 1999 *SAR QSAR Environ. Res.* **10** 545
33. Ponec R, Amat LI and Carbó-Dorca R 1999 *J. Comput.-Aid. Mol. Des.* **13** 259
34. Ponec R, Amat LI and Carbó-Dorca R 1999 *J. Phys. Org. Chem.* **12** 447
35. Amat LI, Carbó-Dorca R and Ponec R 1999 *J. Med. Chem.* **42** 5169
36. Solà M, Mestres J, Oliva JM, Duran M and Carbó R 1996 *Int. J. Quantum Chem.* **58** 361
37. Carbó R, Arnau J and Leyda L 1980 *Int. J. Quantum Chem.* **17** 1185
38. Bultinck P and Carbó-Dorca R 2003 *J. Chem. Inf. Comput. Sci.* **43** 170
39. Constans P, Amat LI and Carbó-Dorca R 1997 *J. Comput. Chem.* **18** 826
40. Bultinck P, Kuppens T, Gironés X and Carbó-Dorca R 2003 3D QSAR modeling in drug design. *J. Chem. Inf. Comput. Sci.* **43** 1143
41. Oprea TI 2004 *Computational medicinal chemistry for drug discovery* (eds) P Bultinck, H De Winter, W Langenaeker and J P Tollenaere (New York: Dekker) pp 571–616
42. Mezey P G, Ponec R, Amat LI and Carbó-Dorca R 1999 *Enantiomer* **4** 371
43. Parr R G and Bartolotti L J 1983 *J. Phys. Chem.* **87** 2810
44. Ayers P W 2000 *Proc. Natl. Acad. Sci. USA* **97** 1959
45. Bultinck P and Carbó-Dorca R 2004 *J. Math. Chem.* **36** 191
46. Hodgkin E E and Richards W G 1987 *Int. J. Quant. Chem. Quantum Biol. Symp.* **14** 105
47. Hodgkin E E and Richards W G 1988 *Chem. Ber.* **24** 1141
48. Parr R G and Yang W 1984 *J. Am. Chem. Soc.* **106** 4049
49. Perdew J P, Parr R G, Levy M and Balduz J L Jr 1982 *Phys. Rev. Lett.* **49** 1691
50. Ayers P W and Levy M 2000 *Theor. Chem. Acc.* **103** 353
51. Geerlings P, Boon G, Van Alsenoy C and De Proft F 2005 *Int. J. Quantum Chem.* **101** 722
52. Yang W and Parr R G 1985 *Proc. Natl. Acad. Sci. USA* **82** 6723
53. Cramer R D III, Patterson D E and Bunce J D 1988 *J. Am. Chem. Soc.* **110** 5969
54. Wagener M, Sadowski J and Gasteiger J 1995 *J. Am. Chem. Soc.* **117** 7769
55. Robert D, Amat LI and Carbó-Dorca R 1999 *J. Chem. Inf. Comp. Sci.* **39** 333
56. Lobato M, Amat LI, Besalú E and Carbó-Dorca R 1997 *Quantum Struct.-Act. Relat.* **16** 465
57. Stewart J J P 1990 Semiempirical molecular orbital methods. *Reviews in computational chemistry* (eds) K B Lipkowitz and D B Boyd (New York: Wiley) vol 1, pp 45–81
58. Bredow T 2004 Semi-empirical methods. *Computational medicinal chemistry for drug discovery* (eds) P Bultinck, H De Winter, W Langenaeker and J P Tollenaere (New York: Dekker) pp 29–55
59. Becke A D 1993 *J. Chem. Phys.* **98** 5648
60. Lee C T, Yang W T and Parr R G 1988 *Phys. Rev.* **B37** 785
61. Stephens P J, Devlin J F and Chabalowski C F 1994 *J. Phys. Chem.* **98** 11623
62. Frisch M J *et al* 2004 Gaussian 03 Revision B.05 (Wallingford, CT: Gaussian Inc.)
63. Helgaker T, Jorgens P and Olsen J 2000 *Molecular electronic structure theory* (New York: Wiley)
64. Gironés X, Robert D and Carbó-Dorca R 2001 *J. Comput. Chem.* **22** 255
65. Maggiora G M, Petke J D and Mestres J 2002 *J. Math. Chem.* **31** 251
66. Carbó R, Besalú E, Amat LI and Fradera X 1996 *J. Math. Chem.* **19** 47
67. Carbó-Dorca R, Besalú E, Amat LI and Fradera X 1996 *Advances in molecular similarity* (eds) R Carbó-Dorca and P G Mezey (London: JAI Press) vol 1, pp 1–42
68. Willett P, Barnard J M and Downs G M 1998 *J. Chem. Inf. Comput. Sci.* **38** 983