

# Molecular Quantum Similarity Matrix Based Clustering of Molecules Using Dendrograms

Patrick Bultinck<sup>\*,†</sup> and Ramon Carbó-Dorca<sup>‡</sup>

Department of Inorganic and Physical Chemistry, Ghent University, Krijgslaan 281 (S-3), B-9000 Gent, Belgium, and Institute of Computational Chemistry, University of Girona, Campus Montilivi, 17071 Girona, Catalonia, Spain

Received September 13, 2002

A new scheme for general classification of quantum objects is presented. Based on molecular quantum similarity matrices (MQSM), different algorithms are presented for generating Molecular Quantum Similarity Dendrograms (MQSD). An application of MQSD is presented for a set of steroid molecules.

## INTRODUCTION

Chemistry is full of instances where molecules are compared to each other to rationalize observations such as chemical reactivity itself and their changes among a set of molecules. The implicit assumption of the quantum similarity (QS) background is such that similar molecules possess similar properties. Although sentences related to molecular comparison are ubiquitous in chemical literature, the question about how similarity between different molecules should be assessed is not yet completely answered. Many attempts to answer this question have been published, and detailed accounts of these attempts may be found in refs 1 and 2 and in ref 3 as well. Some methods to assess molecular similarity are based on the presence or absence of certain features in the two compared molecules. Other methods are based on geometrical considerations. For instance, when evaluating the similarity of conformations of a molecule, the similarity may be assessed via the Euclidean distances between the values of corresponding dihedral angles in the different conformations. Still other methods use similarity in molecular descriptors between different molecules. Many alternative methods have been developed and have been reviewed.<sup>1–3</sup>

Despite all these efforts, there still is no definite answer on how to express or evaluate molecular similarity. From a quantum chemical point of view, the most conclusive method to answer this question is using as a comparison tool the molecular electron density. According to quantum theory electron density is the container of all the information about the properties of a given molecule, and as such electron density comparison is the most promising method to assess similarity. This is the subject of the general point of view where QS is based, as advocated by Carbó et al.<sup>4</sup> in 1980 and deeply developed hereafter.<sup>5</sup> The present study will use the ideas of QS in order to classify molecular structures or their conformations and isomeric forms as well.

Molecular similarity is not only an interesting and important topic in itself but also a cornerstone of many

important research areas in chemistry. Fields such as structure and substructure searching rely heavily on the ideas of molecular similarity. Molecular Quantum Similarity (MQS) itself provides the main basis to the field of quantum QSAR.<sup>6</sup>

Although several molecules can be simultaneously compared,<sup>7–9</sup> similarity is usually calculated for one molecule against another molecule. For a set of molecules, one may then construct a similarity matrix, which allows the identification of the most similar pairs. It is less straightforward to deduce a clustering or grouping of molecules from the MQS matrix (MQSM). The aim of the present study is to develop a classification methodology for quantum objects in general, starting from a MQSM. This classification should help the identification of groups or clusters of quantum objects versus other groups of quantum objects. Also, the scheme should be applicable in principle over all kinds of quantum objects that is, sets of molecules from congeneric or noncongeneric series as well as sets of different conformations of the same molecule. In fact QS has been employed in the study of nuclear structure and classification.<sup>10,11</sup>

## THEORETICAL BACKGROUND

The following paragraphs intend to introduce the reader into some of the basic elements of MQS. A complete review of all theoretical features is well out of the scope of the present work. The complete theory was reviewed previously by Carbó et al. in several papers.<sup>12–15</sup>

Molecules are obvious examples of so-called quantum objects. A quantum object is defined as an element of a collection of composite mathematical entities made essentially by two ordered parts: submicroscopic systems and their tags the agglomerated quantum probability density functions.<sup>16–18</sup> Specifically for molecules, such a quantum object has as its object part the identity of the molecular structure in a Born–Oppenheimer approximate way and as tags the electron density functions. For the present derivations, one can identify the identity of the molecule most easily with the number or name of the molecule in the set of molecules, with the attachable set of nuclear identities and their coordinates. The tag of a given molecular identity can be associated with the electron density  $\rho(r)$  of the

\* Corresponding author phone: +32/9/264.44.23; fax: +32/9/264.49.83; e-mail: Patrick.Bultinck@rug.ac.be.

<sup>†</sup> Ghent University.

<sup>‡</sup> University of Girona.

molecule. Once the set of quantum objects for the study at hand is defined, it is fairly straightforward to introduce a quantum similarity measure (QSM),<sup>9,19,20</sup> involving two or more quantum objects. Suppose there are two quantum objects, one with the identity part A and the other one with identity B. Their tags will be chosen as the electron densities  $\rho_A(\mathbf{r})$  and  $\rho_B(\mathbf{r})$ . Electron densities are positive definite everywhere in the appropriate domain of definition, and then with a positive definite operator  $\Omega$ , one can introduce the QSM as follows:

$$Z_{AB} = \langle \rho_A | \Omega | \rho_B \rangle = \int \int \rho_A(\mathbf{r}_1) \Omega(\mathbf{r}_1, \mathbf{r}_2) \rho_B(\mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2 \quad (1)$$

The integrations are carried out over a proper domain or, as usual, over the entire space where molecular structures are supposed to be embedded. The operator appearing in eq 1 may be chosen over several possibilities, including Dirac delta function, Coulomb operator,<sup>21–23</sup> or kinetic energy.<sup>24,25</sup> In the present study a Dirac delta function will be used, i.e.,  $\Omega(\mathbf{r}_1, \mathbf{r}_2) \rightarrow \delta(\mathbf{r}_1 - \mathbf{r}_2)$  so that the integral in eq 1 becomes similar to a well-known overlap integral involving two density functions. When calling  $Z_{AB}$  such an overlap QSM, it is obtained from the general definition above:

$$Z_{AB} = \langle \rho_A | \rho_B \rangle = \int \rho_A(\mathbf{r}) \rho_B(\mathbf{r}) d\mathbf{r} \quad (2)$$

Since electron densities are positive definite and the Dirac delta operator is a positive definite operator too, the overlap QSM become always positive definite real numbers, that is

$$Z_{AB} \in \mathbf{R}^+ \quad (3)$$

Supposing a set of  $N$  quantum objects, it is easy to calculate all QSM for each combination of available quantum objects pairs. This yields a MQS matrix (MQSM):  $\mathbf{Z}$ , with elements defined as in eq 4:

$$\mathbf{Z} = \{Z_{IJ}\} = \begin{bmatrix} Z_{11} & Z_{12} & \dots & Z_{1N} \\ Z_{21} & Z_{22} & \vdots & Z_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{N1} & Z_{N2} & \dots & Z_{NN} \end{bmatrix} \quad (4)$$

The different matrix elements in the  $\mathbf{Z}$  matrix can be computed in general by formulas like eq 2, and as such the matrix  $\mathbf{Z}$  is a positive definite matrix, provided that the implied integrals are computed by means of the same set of nuclear coordinates for every quantum object. The diagonal elements,  $Z_{AA}$ , are called self-similarity measures and are themselves often useful QSAR descriptors,<sup>23,26,27</sup> related to classical concepts as Hammett  $\sigma$  or octanol–water partition coefficients.

The most interesting features of the matrix  $\mathbf{Z}$  are that for every quantum object of the set of  $N$  objects, new discrete tags can be introduced instead of the electron density ones. These discrete tags correspond to the elements of the  $\mathbf{Z}$  column vector associated to each quantum object. The quantum object A, for example, is defined within the quantum object set through its elements  $Z_{KA} (K = 1, N)$ . This replaces the spatially smeared out tag  $\rho$  with an  $N$ -dimensional vector of discrete real positive definite values. This results in some important properties of these descriptors, namely the fact that they are as follows:<sup>28,29</sup> (1) universal, since they can be obtained for any quantum object and (2) unbiased, since they

do not involve any other choice than the operator used (Dirac operator in the present study) in evaluating the elements of the MQSM  $\mathbf{Z}$ .

The matrix  $\mathbf{Z}$  and its unique properties can be now used for the exploration of a way to classify quantum objects according to their tags.

## DENDROGRAM CONSTRUCTION

There exist a large number of methods to cluster quantum objects. Examples are neural networks, statistical techniques such as PCA, and many others. In the present study we wish to use a technique which does not introduce new concepts that cannot always be directly related to  $\mathbf{Z}$ . Furthermore, the method can be chosen hierarchical and unbiased in the number of factors used in statistical techniques. As will be shown, dendrograms, which fulfill the previous properties, are quite an interesting and easily implemented method from the computational point of view. When dendrograms are chosen as a tool, a further decision has to be taken, namely the use of an agglomerative or a divisive scheme. In an agglomerative scheme, starting from the set of parent quantum objects, subsequent steps involve grouping of different, individual or groups of quantum objects in a larger collection. The divisive schemes involve splitting of groups of quantum objects in individual quantum objects or smaller groups of quantum objects, starting from the cluster holding all parent quantum objects. The notion of parent quantum objects refers to the original quantum objects, in this case the individual molecules in the set. These quantum objects are those that compose the original  $\mathbf{Z}$  matrix. Both schemes of choice end when no more agglomerations or divisions can be performed, that is when all quantum objects are gathered in one group, or no more divisive steps can be taken, and thus every quantum object stands individually in the dendrogram. It has been shown that for many applications, agglomerative schemes are highly preferable, since they tend to yield much fewer cases of dubious decision-making with respect the question of where to classify an object.<sup>30</sup>

It is clear from eqs 2 and 4 that the elements of the MQSM  $\mathbf{Z}$  only contain direct information relating one molecule to any other molecule. As such, only very limited information for clustering the molecules is directly available from  $\mathbf{Z}$ . A further problem consists in that the elements of  $\mathbf{Z}$  are positive definite but not restricted in magnitude. They have no upper bounds. This makes it difficult to use the values of the  $\mathbf{Z}$  matrix elements in numerical comparisons. Such kinds of problems will be solved in different ways, as will be discussed below.

**Dendrograms using Carbó Similarity Indices (CSI).** An elegant way to solve the previously mentioned problems consists of using Carbó similarity indices (CSI).<sup>31</sup> All elements of the matrix  $\mathbf{Z}$  can then be transformed in a number lying in the interval (0,1]. This transformation involves the calculation of the elements  $C_{AB}$  of the matrix  $\mathbf{C}$  through

$$C_{AB} = \frac{Z_{AB}}{\sqrt{Z_{AA}Z_{BB}}} \quad (5)$$

It is clear that the diagonal elements of the matrix  $\mathbf{C}$ , obtained from the diagonal elements of  $\mathbf{Z}$  are transformed into one, and all other elements will lie within the interval

(0,1]. This can be interpreted as that the extent of the deviation from 1 of an element  $C_{AB}$  reflects the extent of dissimilarity between molecules A and B. The construction of the agglomerative dendrogram then shall proceed by searching the sequences of largest off-diagonal elements of **C**.

However, once such a first step is taken, and two quantum objects have been gathered in one new cluster, the matrix **Z** is no longer applicable as it is. Consider that in a first step molecules A and B have been gathered, then all matrix elements  $Z_{KA}$  and  $Z_{KB}$  ( $K = 1, N$ ) have lost their ordering meaning. Instead, to obviate such a problem, a new object is introduced. This new object is, in fact, the cluster consisting of elements A and B and is a quantum object itself. Although it can mathematically be treated as a quantum object, it is not a physical object like a molecule but is an artificial object. It is a quantum object since it may be given an identity as "cluster 1" and a tag. This tag has to be obtained in some way from the matrix  $\mathbf{Z}^0$ , where the superscript  $^0$  denotes the original startup matrix **Z**. For each of the  $N$  discrete descriptors in the column vector for the new quantum object, the average is taken of the two quantum objects composing this new quantum object. Denoting A and B the quantum objects gathered in the new quantum object X, one has

$$\forall K \in \{1, \dots, N\}: Z_{KX} = \frac{Z_{KA}^0 + Z_{KB}^0}{2} \quad (6)$$

Note that for the consistency of the **Z** and **C** definitions, it should be stressed that averages are made in the elements of **Z** and *not* in the elements of **C**. Using the above averaging method, the matrix **Z** remains symmetric. This is so because averages are convex combinations of the  $\mathbf{Z}^0$  matrix columns, and as such the result can be always considered a new quantum object.

Once the first cluster has been constructed, holding elements A and B, the new matrix **Z** is constructed. From this matrix, a new matrix **C** may be constructed, where the terms in the denominator of eq 5 are taken from the new **Z** matrix. The process of constructing new **Z** matrices and searching the biggest off-diagonal element of **C** may be repeated unchanged at every stage in the construction of the dendrogram. Since every new agglomeration into a new averaged quantum object involves two quantum objects A and B, one could opt to calculate an average of the elements of the column vectors of both quantum objects A and B. This would, however, continuously reduce the weight of the first two associated objects in the subsequent clustering steps, which is clearly not intended. Therefore, after every stage in the dendrogram construction, all elements of the **Z** matrix are reconstructed as

$$Z_{XY} = \frac{1}{N_X} \frac{1}{N_Y} \sum_{I=1}^{N_X} \sum_{J=1}^{N_Y} Z_{IJ}^0 \quad (7)$$

$I \in X \quad J \in Y$

where  $N_X$  is the number of parent quantum objects present in the cluster X and  $N_Y$  is the number of parent objects present in cluster Y. The so-called parent quantum objects are hereby those individual quantum objects that form the matrix  $\mathbf{Z}^0$ . The indices  $I$  and  $J$  refer to the parent quantum objects

contained in the clusters X and Y, respectively. Equation 7 in fact shows that the technique used here is a Sequential Agglomerative Hierarchical Nonoverlapping (SAHN) method.<sup>30</sup> The scheme for updating the matrix through eq 7 corresponds to the so-called Group average.<sup>30</sup>

Computationally the procedure starts from the matrix  $\mathbf{Z}^0$ , which is transformed immediately into the matrix **Z** using eq 7 for every element (where in the first step  $N_X = N_Y = 1$ ). From this new similarity matrix, matrix **C** is formed via eq 5. The highest nondiagonal element of **C** is used to decide which two objects X and Y are joined in a new quantum object. For each element of **Z** involving the joined elements, the new values  $Z_{(XY),K}$  ( $K = 1, N$ ) are calculated from eq 7, and a new **Z** matrix is constructed. This entire procedure is repeated until all parent objects have been agglomerated in the ultimate all-agglomerative object. Although the number of objects is reduced continuously as the dendrogram is constructed, the implementation of the algorithm keeps on working with  $(N \times N)$  **Z** and **C** matrices, where the averages from eq 7 are stored in the column vector of the first element of the agglomeration (A). The column and row vectors of the second element (B) are filled with zero elements, except for the diagonal  $Z_{BB}$  elements which are given a value of one, to avoid infinite numbers in eq 5. This method strongly reduces the computational load, compared to having to redimension matrices and shift matrix elements.

**Dendrograms using a Stochastic Transformation.** A second way to manipulate the **Z** matrix to solve the problem of undetermined positive definite intervals in **Z** consists of using stochastic transformations.<sup>32</sup> This involves constructing a new matrix **S**, by means of

$$S_{IA} = Z_{IA} \left( \sum_{K=1}^N Z_{KA} \right)^{-1} \quad (I = 1, N) \quad (8)$$

This means that every element in a given column A of the matrix **Z** is divided by the sum of all the elements of the chosen column A. This immediately entails that the sum over all elements in a column A,  $|S_A\rangle$  of matrix **S** is equal to unity, or expressed in short by  $\langle |S_A\rangle = 1$ . As a result the matrix **S**, contrary to the original MQSM **Z**, is not symmetric.

Proceeding from this stochastic matrix **S**, different options may be taken to start constructing the dendrogram. The most intuitive one lies in using the column vectors as coefficients to obtain a symmetrical matrix **C** as

$$\mathbf{C} = \mathbf{S}^T \mathbf{S} \quad (9)$$

Intuitively, the most similar set of two quantum objects A and B would then correspond to the biggest matrix element  $C_{AB}$  of **C**. Test calculations revealed that in most cases for a set of molecules with quantum self-similarity measures of about the same magnitude, this approach works well. However, in some cases, it was found that when the same molecule was entered twice in the quantum object set, depending on the column vectors of this molecule, on some occasions it was not identified as the most similar couple, which clearly becomes a flaw. This misbehavior can occur under certain conditions, which are not easily checked in the early stages of the dendrogram construction. This approach, however, with intuitive background was therefore abandoned.



A possible solution of this above-mentioned problem is to subtract from each element  $C_{AB}$  in a column the diagonal element  $C_{BB}$  and take the absolute value of this difference afterward. In such case, the smallest element of **C** indicates the most similar pair. Unfortunately, the resulting matrix **C** is not symmetrical. Although found to work properly for the examples tested, this unsymmetrical feature is conditioned to reject such an approach.

A different approach to retaining the symmetry in **C** is based on distance measures between the column vectors of **S**. Column  $|S_A\rangle$  of matrix **S** identifies the quantum object *A* within the quantum object set. A similar argument holds for  $|S_B\rangle$ . If now, all components are very similar this means that the position of objects *A* and *B* in the *N*-dimensional space is very similar, indicating high molecular similarity. One can then introduce a Minkowski metric between quantum objects *A* and *B*. In general, Minkowski metrics for  $C_{AB}$  are given by

$$C_{AB} = \sqrt[r]{\sum_{K=1}^N |S_{KA} - S_{KB}|^r} \quad (10)$$

using  $r = 1$ , this gives the Manhattan distance, in this case for  $C_{AB}$

$$C_{AB} = \sum_{K=1}^N |S_{KA} - S_{KB}| \quad (11)$$

or alternatively, choosing  $r = 2$  an Euclidean distance is defined:

$$C_{AB} = \sqrt{\sum_{K=1}^N |S_{KA} - S_{KB}|^2} \quad (12)$$

Both algorithms conduct to well-defined distance measures.<sup>31</sup> Using such measures, the shortest distance, and thus the lowest off-diagonal element of **C**, indicates the most similar pair of quantum objects. Once both objects have been identified, one can construct a new matrix **Z**, as was described above, see also eq 7. Using this new **Z**, a new stochastic transformation may be made, and a new **C** is obtained to identify the next most similar couple of quantum objects. This procedure is then repeated until all parent quantum objects have been gathered in one single quantum object.

All possible methods above-discussed have been implemented and checked for their validity. It was found that in nearly all cases, even the less sound method, represented by the matrix algorithm of eq 9, produces reasonable results. Due to, probably seldom, but possible failures, the approach based on eq 9 is excluded for the application presented below.

#### DENDROGRAM CONSTRUCTION FOR A SET OF GLOBULIN BINDING STEROIDS

To test the classification approach presented above, a set of 31 globulin binding steroids, as used previously by Cramer et al.,<sup>33</sup> Wagener et al.,<sup>34</sup> and Carbó et al.,<sup>35,36</sup> was used. The structures and names of the steroids are given in Table 1.

Geometries for the different molecules were obtained using AM1 geometry optimizations.<sup>37</sup> The **Z** matrix for the entire object set was constructed using the Atomic Shell Approximation calculated electron densities.<sup>38–41</sup> Although this standard approach was employed here, the schemes developed above are independent of the method used to generate the density tags for the quantum objects.

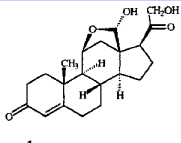
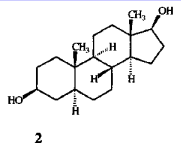
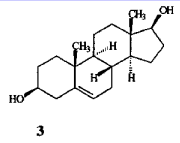
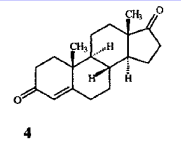
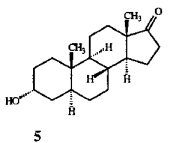
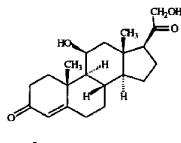
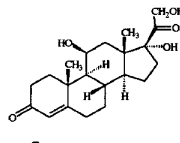
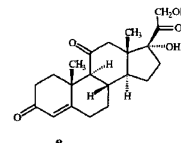
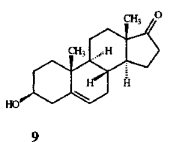
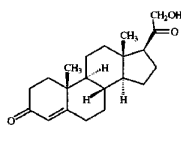
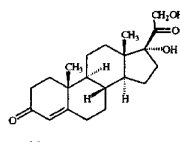
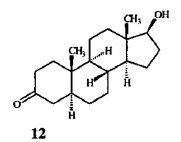
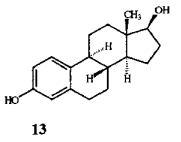
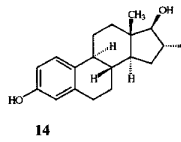
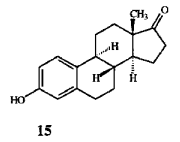
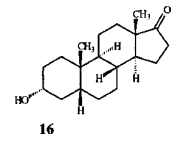
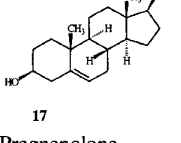
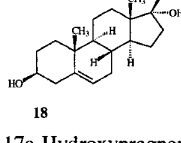
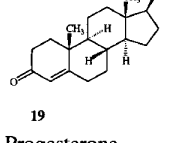
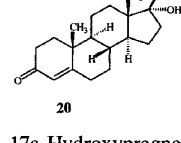
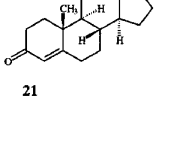
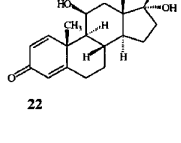
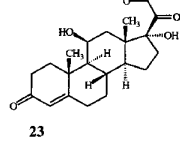
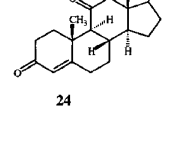
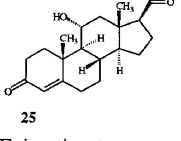
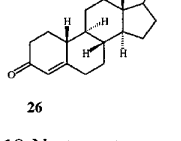
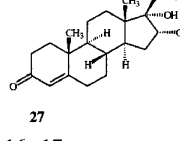
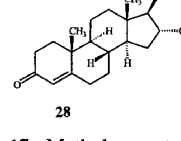
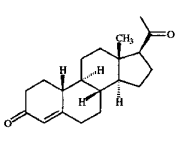
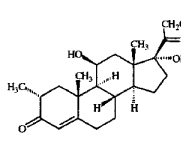
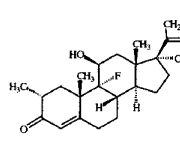
For all schemes developed above, the flowchart for program execution is quite simple and is presented in Figure 1. In the case of the Carbó similarity index scheme, no **S** matrix is introduced, but rather the **C** matrix is calculated directly from **Z**.

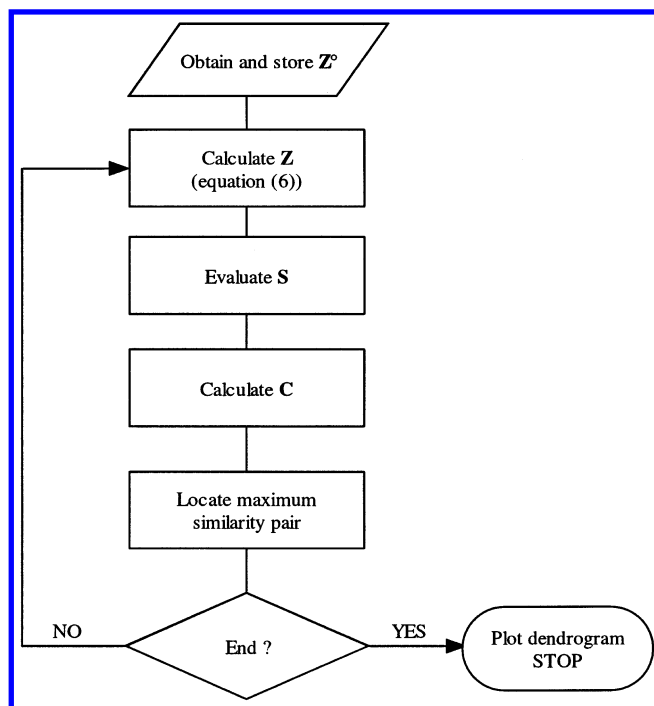
The “END ?” decision step involves checking whether all parent quantum objects have been gathered in one single final cluster. This is the case after *N*-1 clusters have been constructed, where *N* is the number of parent quantum objects. The criterion to decide perfect similarity depends on the scheme used. In case of the Carbó similarity index, the highest similarity is indicated by the largest value of an off-diagonal element. In case of the distance measures in eq 11 or 12, molecules are most similar for the lesser off-diagonal element. Even without an attempt to construct the computer implementations as highly performing as possible, the programs succeed at clustering the 31 steroid molecules in less than 0.5 s on a medium range personal computer.

The Carbó similarity index scheme yielded the results shown in Figure 2.

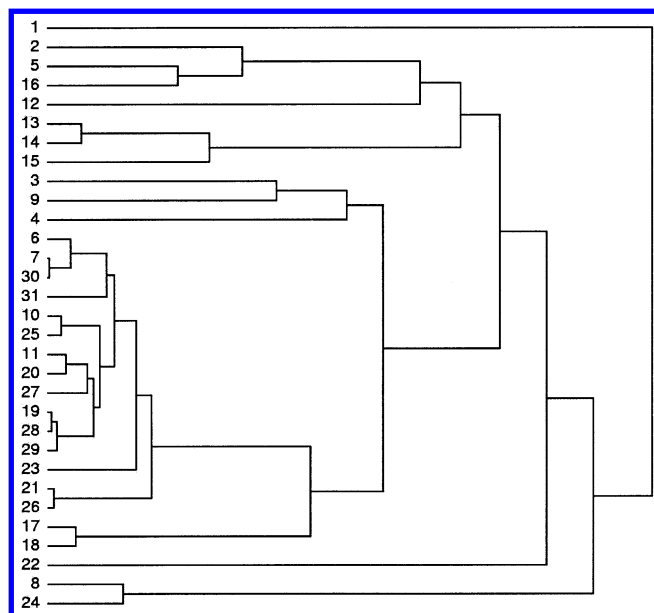
A complete discussion of the clustering is out of the context of the present study, but several interesting features can be noted. Furthermore, a few relatively easy instances where the similarity is relatively obvious may be identified from direct inspection of the molecular structures. The dendrograms should reflect this high logic-visual similarity. A very obvious case lies in molecules 7 and 30. These have very similar AM1 optimal geometries, and only differ in the presence of a methyl group in some molecular position. The original **Z**<sup>o</sup> matrix also reflects this high similarity, and in fact it is the first cluster that is identified in the dendrogram as well. The Carbó similarity index between this molecular pair is 0.98568, which indicates such high similarity. A similar case exists between molecules 19 and 28, where the only difference is again a single methyl group, and again a very early clustering is found ( $C_{AB} = 0.98100$ ). The third cluster is again an illustration of the same effect. These clusterings are quite easy to detect from simple inspection of the **Z**<sup>o</sup> matrix. The more interesting cases arise when the dendrogram construction procedure starts combining groups of quantum objects with other groups or individual parent quantum objects. This is the case for the fourth cluster. Although molecules 10 and 25 are also highly similar one to each other, differing in one hydroxyl group this time, the clustering indicates that there is a higher similarity between the quantum object built from molecules 19 and 28 on one hand and molecule 29 on the other hand. This would not have been seen from the **Z**<sup>o</sup> matrix, which indicates a clear advantage of the present dendrogram. Molecules 29, 19, and 28 all have the same basic framework, in fact the entire molecule 29, with the addition of a single methyl group in 19, and a second one in molecule 28. The 11th cluster is the first where quantum objects, which are clusters themselves, are joined. That this cluster would be created could not be predicted from simple chemical reasoning, since the mol-

**Table 1.** Molecular Structures and Names for the 31 Molecules Contained in the Steroid Set

			
1 Aldosterone	2 Androstenediol	3 5-Androstenediol	4 4-Androstenedione
			
5 Androsterone	6 Corticosterone	7 Cortisol	8 Cortisone
			
9 Dehydroepiandrosterone	10 11-deoxycorticosterone	11 11-deoxycortisol	12 Dihydrotestosterone
			
13 Estradiol	14 Estrinol	15 Estrone	16 Ethiochalconone
			
17 Pregnenolone	18 17a-Hydroxypregnenolone	19 Progesterone	20 17a-Hydroxypregnenolone
			
21 Testosterone	22 Prednisolone	23 Cortisolacetat	24 4-Pregnene-3,11,20-trione
			
25 Epicorticosterone	26 19-Nortestosterone	27 16a,17a-Dihydroxyprogesterone	28 17a-Methylprogesterone
			
29 19-Norprogesterone	30 2a-Methylcortisol	31 2a-Methyl-9a-Fluorocortisol	

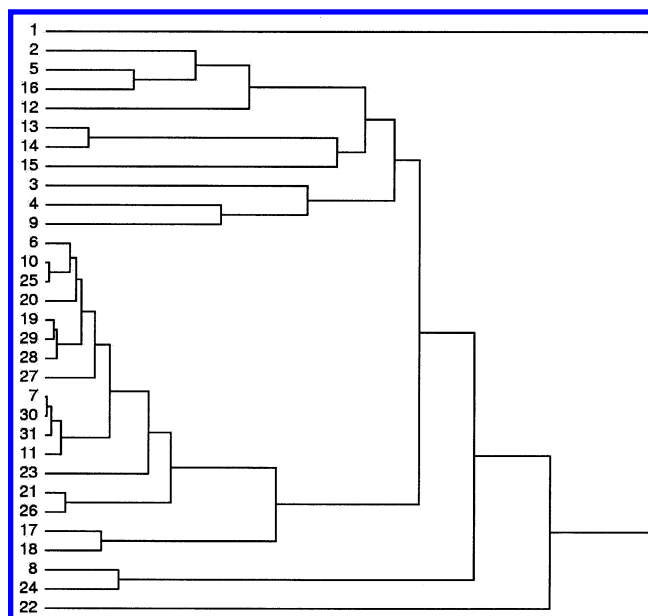


**Figure 1.** General flowchart for the construction of the dendrograms.



**Figure 2.** Dendrogram obtained using the Carbó similarity index based algorithm for the set of 31 steroids.

ecules start differing by much more than a simple methyl group. Another interesting observation lies in the fact that some molecules remain outliers of any cluster for a very long time along the clustering procedure. The most prominent example is molecule 1, which is an outlier for every cluster, except in the ultimate stage of the process, where by construction it has to be taken in a conclusive cluster. For this specific molecule, which contains an extra ring, chemical intuition may have predicted this. Another example is molecule 22, which is also isolated for a very long set of steps, but for which it is not so straightforward to predict that it can be an isolated structure. Also interesting is the fact consisting in that during the entire procedure, even in later stages, cases are found where rather than the little-

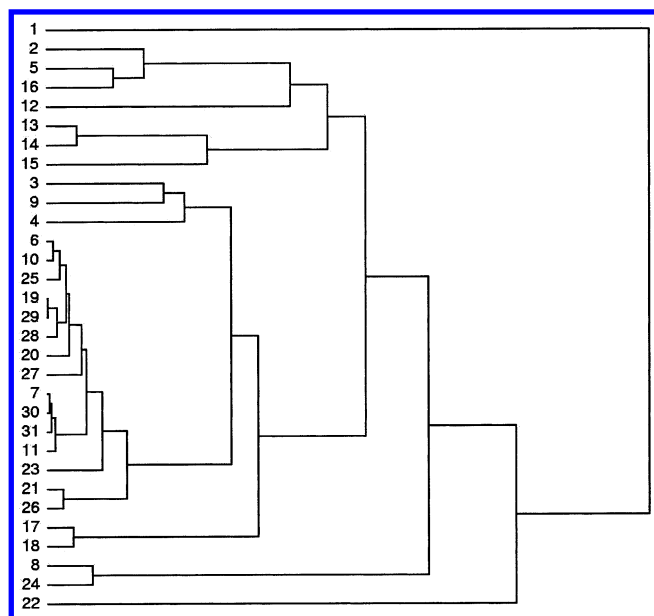


**Figure 3.** Dendrogram obtained using the stochastic transform and Manhattan distance algorithm for the set of 31 steroids.

informative scenario of clusters picking up an extra molecule, completely new clusters are formed from two parent quantum objects. Clusters 18 and 21 are fine examples of this. Concerning the values of the maximum Carbó similarity index used to identify new clusters, it is found that during the entire procedure, they usually decrease. For the first cluster the maximum Carbó similarity index is 0.98568, for the last one it is 0.47273. Over the entire clustering procedure, there is nearly a continuous decrease, although some cases occur where there is a slight increase. This is not a flaw, since it should be kept in mind that whenever two quantum objects are joined to form a new one, for all remaining quantum objects their column vectors are changed according to eq 7. This is equivalent to realizing that after every new creation of a quantum object, a new  $\mathbf{Z}$  is constructed.

Figures 3 and 4 give the dendrograms obtained using the stochastic-Manhattan and stochastic-Euclidean schemes, respectively.

Looking at the Manhattan scheme, one finds that the closest cluster, the one identified first, is again composed of molecules 7 and 30. Also very similar with the Carbó similarity index dendrogram appears the fact by which molecule 1 remains an outlier until the creation of the very last cluster. Between these two extremes there are, however, differences between Manhattan and the Carbó similarity index dendrograms. Usually these are only minor differences, and these are concentrated in the earlier stages of the dendrogram. There, for the present set of molecules, the decision coordinate, the maximum or minimum element of  $\mathbf{C}$ , differs only a little between different steps. Minor differences between  $\mathbf{C}$  for both schemes can, as such, cause some reversals. This is the case for Carbó similarity index cluster 5 and Manhattan cluster 2. In both cases molecules 10 and 15 are joined but at a somewhat earlier stage in case of Carbó similarity index. Looking at Carbó similarity index cluster 16 and Manhattan cluster 17, one finds that they are composed of completely the same molecules, although the clusters were grown in a somewhat different way. When



**Figure 4.** Dendrogram obtained using the stochastic transform and Euclidean distance algorithm for the set of 31 steroids.

proceeding, the same molecule 26 is added, and the final clustering structure is very similar. Concerning the evolution of the smallest C element, very similar behavior was found as in case of the Carbó similarity index method.

An in-depth discussion of the dendrogram obtained, using both stochastic matrices and Euclidean distances, is not given, again due to the high similarity of the dendrogram with the previously discussed ones. It is clear that the dendrograms are all quite similar, appearing independent of the dendrogram algorithm used. This is clearly advantageous since it seems that the choice of one construction scheme or another alternative does not introduce a bias influencing the results.

### CONCLUSIONS

A new technique for quantum object sets clustering, using molecular quantum similarity measures is introduced. Using overlap integrals between the tags of the set of quantum objects, the molecular quantum similarity matrix is constructed. This matrix was then transformed using Carbó similarity index, and as such employed to construct dendrograms, which hierarchically cluster the involved quantum objects in the set. At each stage, new averaged quantum objects are introduced, and new quantum similarity matrices constructed, which in turn are used to cluster the original quantum objects. Other schemes involve the use of stochastic matrices obtained from the molecular similarity matrices and the use of different distance measures to identify the most similar quantum objects.

The proposed schemes were tested on a set of steroid molecules, and were found to give much extra information on the clustering of the compounds. This information is not readily visible from the original molecular quantum similarity matrix, enhancing the information contained in the dendrograms.

The present methodology is generally applicable to any set of quantum objects. As a consequence, they can be used equally well for the study of similarity and clustering of different molecules, as well as for single molecule conformational clustering.

### ACKNOWLEDGMENT

P. Bultinck wishes to thank the *Fund for Scientific Research-Flanders* (Belgium) for their grants to the Computational Chemistry group at Ghent University and acknowledges the *European Community – Access to Research Infrastructure action of the Improving Human Potential Program*, allowing the use of the CEPBA infrastructure at the PolyTechnical University of Catalonia (Spain) and the fellowship with the Institute of Computational Chemistry at the University of Girona (Catalonia, Spain). R. Carbó-Dorca acknowledges the Foundation M. F. de Roviralta as well as the CICYT project #SAF2000-223, which have supported this work. The authors wish to deeply acknowledge Dr. X. Gironés of the Institute of Computational Chemistry, University of Girona, for kindly providing the quantum similarity matrices, employed for the reported application examples in this study. Dr. Peter Kleiweg is acknowledged for providing the postscript dendrogram drawing program “den” (<http://odur.let.rug.nl/~kleiweg/clustering/clustering.html>).

### REFERENCES AND NOTES

- (1) *Molecular Similarity in Drug Design*; Dean, P. M., Ed.; Blackie Academic & Professional: New York, 1995.
- (2) *Concepts and applications of molecular similarity*; Johnson, M. A., Maggiora, G. M., Eds.; Wiley-Interscience: New York, 1990.
- (3) *Fundamentals of Molecular Similarity*; Carbó-Dorca, R., Gironés, X., Mezey, P. G., Eds.; Kluwer Academic/Plenum Publishers: New York, 2001.
- (4) Carbó, R.; Leyda, L.; Arnau, M. How Similar is a Molecule to Another? An Electron Density Measure of Similarity between Two Molecular Structures. *Int. J. Quantum Chem.* **1980**, *17*, 1185–1189.
- (5) Carbó-Dorca, R.; Amat, L.; Besalú, E.; Gironés, X.; Robert, D. In *Fundamentals of Molecular Similarity*; Carbó-Dorca, R., Gironés, X., Mezey, P. G., Eds.; Kluwer Academic/Plenum Publishers: New York, 2001; pp 187–191.
- (6) Carbo-Dorca, R.; Amat, L.; Besalu, E.; Gironés, X.; Robert, D. Quantum Mechanical Origin of QSAR: Theory and Applications. *J. Mol. Struct. (THEOCHEM)* **2000**, *504*, 181–228.
- (7) Robert, D.; Carbó-Dorca, R. Analyzing the triple density molecular quantum similarity measures with the INDSCAL model. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 620–623.
- (8) Carbó, R.; Calabuig, B.; Besalú, E.; Martínez, A. Triple Density Molecular Quantum Similarity Measures: A General Connection Between Theoretical Calculations and Experimental Results. *Mol. Eng.* **1992**, *2*, 43–64.
- (9) Carbó-Dorca, R.; Amat, L.; Besalú, E.; Lobat, M. In *Advances in Molecular Similarity*; Carbo-Dorca, R., Mezey, P. G., Eds.; JAI Press: Stanford, CT, 1998; pp 1–42.
- (10) Robert, D.; Carbó-Dorca, R. On the extension of QS to atomic nuclei: Nuclear QS. *J. Math. Chem.* **1998**, *23*, 327–351.
- (11) Robert, D.; Carbó-Dorca, R. Structure–property relationships in nuclei. Prediction of the binding energy per nucleon using a quantum similarity approach. *Il Nuovo Cimento* **1998**, *A111*, 1311–1321.
- (12) Carbó-Dorca, R.; Besalú, E. A general survey of Molecular Quantum Similarity. *J. Mol. Struct. (THEOCHEM)* **1998**, *451*, 11–23.
- (13) Carbó-Dorca, R.; Robert, D.; Amat, L.; Gironés, X.; Besalú, E. Molecular Quantum Similarity in QSAR and Drug Design. *Lecture Notes Chem.* **2000**, *73*.
- (14) Carbó, R.; Besalú, E. In *Molecular similarity and reactivity: From quantum chemical to phenomenological approaches*; Carbó, R., Ed.; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1995; pp 3–30.
- (15) Carbó, R.; Besalú, E.; Amat, L.; Fradera, X. Quantum molecular similarity measures (QMSM) as a natural way leading towards a theoretical foundation of quantitative structure-properties relationship. *J. Math. Chem.* **1995**, *18*, 237–246.
- (16) Carbó-Dorca, R. Tagged sets, convex sets and QS measures. *J. Math. Chem.* **1998**, *23*, 353–364.
- (17) Carbó-Dorca, R. Fuzzy sets and Boolean Tagged sets. *J. Math. Chem.* **1997**, *22*, 143–147.
- (18) Carbó-Dorca, R. In *Advances in Molecular Similarity*; Carbo-Dorca, R., Mezey, P. G., Eds.; JAI Press: Stanford, CT, 1998; pp 43–72.
- (19) Besalú, E.; Carbó, R.; Mestres, J.; Solà, M. *Foundations and Recent Developments of Quantum Molecular Similarity. Topics in current*



- Chemistry: Molecular Similarity I*; Springer-Verlag: Berlin, 1995; Vol. 173, pp 31–62.
- (20) Carbó, R.; Besalú, E.; Calabuig, B.; Vera, L. Molecular Quantum Similarity: Theoretical Framework, Ordering Principles and Visualisation Techniques. *Adv. Quantum Chem.* **1994**, 25, 253–313.
- (21) Robert, D.; Gironés, X.; Carbó-Dorca, R. Quantification of the Influence of Single Point Mutations on Haloalkane Dehalogenase Activity: A Molecular Quantum Similarity Study. *J. Chem. Inf. Comput. Sci.* **2000**, 40, 839–846.
- (22) Robert, D.; Amat, L.; Carbó-Dorca, R. Quantum Similarity QSAR: Study of Inhibitors Binding to Thrombin, Trypsin and Factor Xa, Including a Comparison with CoMFA and CoMSIA Methods. *Intl. J. Quantum Chem.* **2000**, 80, 265–282.
- (23) Gironés, X.; Amat, L.; Robert, D.; Carbó-Dorca, R. Use of electron–electron repulsion energy as a molecular descriptor in QSAR and QSPR studies. *J. Comput.-Aided Mol. Des.* **2000**, 14, 477–485.
- (24) Gironés, X.; Gallegos, A.; Carbó-Dorca, R. Modeling Antimalarial Activity: Application of Kinetic Energy Density Quantum Similarity Measures as Descriptors in QSAR. *J. Chem. Inf. Comput. Sci.* **2000**, 40, 1400–1407.
- (25) Carbó-Dorca, R.; Besalú, E.; Gironés, X. Extended density functions. *Adv. Quantum Chem.* **2000**, 20, 3–63.
- (26) Ponec, R.; Amat, L.; Carbó-Dorca, R. Molecular basis of quantitative structure-properties relationships (QSPR): A quantum similarity approach. *J. Comput.-Aided Mol. Des.* **1999**, 13, 259–270.
- (27) Ponec, R.; Amat, L.; Carbó-Dorca, R. Quantum similarity approach to LFER: Substituent and solvent effects on the acidities of carboxylic acids. *J. Phys. Org. Chem.* **1999**, 12, 447–454.
- (28) Carbó-Dorca, R. Inward Matrix Products: Extensions and Applications to Quantum Mechanical Foundations of QSAR. *J. Mol. Struct. (THEOCHEM)* **2001**, 357, 41–54.
- (29) Carbó-Dorca, R. Quantum Quantitative Structure–Activity Relationships (QQSAR): A comprehensive discussion based on Inward Matrix Products, employed as a tool to find approximate solutions of strictly positive linear systems and providing a QSAR-Quantum Similarity Measures. Proceedings of Eccomas 2000 (Barcelona) 2000.
- (30) Sokal, R. R.; Sneath, P. H. A. *Numerical Taxonomy: The Principles and Practice of Numerical Classification*; W. H. Freeman & Co.: London, 1973.
- (31) Carbó R., Besalú E.; Amat L.; Fradera X., On Quantum Molecular Similarity Measures (QMSM) and Indices (QMSI). *J. Math. Chem.* **1996**, 19, 47–56.
- (32) Carbó-Dorca, R. Stochastic Transformation of Quantum Similarity Matrices and Their Use in Quantum QSAR (QQSAR) Models. *Intl. J. Quantum Chem.* **2000**, 79, 163–177.
- (33) Cramer III, R. D.; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect on Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, 110, 5959–5967.
- (34) Wagener, M.; Sadowski, J.; Gasteiger, J. Autocorrelation of Molecular Surface Properties for Modeling Corticosteroid Binding Globulin and Cytosolic Ah Receptor Activity by Neural Networks. *J. Am. Chem. Soc.* **1995**, 117, 7769–7775.
- (35) Robert, D.; Amat, L.; Carbó-Dorca, R. 3D QSAR from tuned molecular quantum similarity measures: Prediction of the CBG binding affinity for a steroids family. *J. Chem. Inf. Comput. Sci.* **1999**, 39, 333–344.
- (36) Lobato, M.; Amat, L.; Besalú, E.; Carbó-Dorca, R. Structure–Activity Relationships of a steroid family using QSM and topological QS Indices. *Quant. Struct.-Act. Relat.* **1997**, 16, 465–472.
- (37) Dewar, M. J. S.; Zebisch, E. G.; Healy, E. F.; Stewart, J. J. P. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, 107, 3902–3909.
- (38) Constans, P.; Carbó, R. Atomic Shell Approximation: Electron Density Fitting Algorithm Restricting Coefficients to Positive Values. *J. Chem. Inf. Comput. Sci.* **1995**, 35, 1046–1053.
- (39) Constans, P.; Amat, L.; Fradera, X.; Carbó-Dorca, R. In *Advances in Molecular Similarity*; Carbó-Dorca, R., Mezey, P. G., Eds.; JAI Press: London, 1996; Vol. 1, pp 187–211.
- (40) Amat, L.; Carbó-Dorca, R. Quantum Similarity Measures under Atomic Shell Approximation: First-Order Density Fitting using Elementary Jacobi Rotations. *J. Comput. Chem.* **1997**, 18, 2023–2039.
- (41) Amat, L.; Carbó-Dorca, R. Fitted Electronic Density Functions from H to Rn for use in Quantum Similarity Measures: Cis-diammine-dichloroplatinum(II) complex as an Application Example. *J. Comput. Chem.* **1999**, 20, 911–920.

CI025602B