Quality of Approximate Electron Densities and Internal Consistency of Molecular Alignment Algorithms in Molecular Quantum Similarity

Patrick Bultinck,*,[†] Ramon Carbó-Dorca,[‡] and Christian Van Alsenoy[§]

Department of Inorganic and Physical Chemistry, Ghent University,
Krijgslaan 281 (S-3), B-9000 Gent, Belgium, Institute of Computational Chemistry, University of Girona,
Campus Montilivi, 17071 Girona, Catalonia, Spain, and Department of Chemistry, Universiteit Antwerpen,
Universiteitsplein 1, B-2610 Antwerpen, Belgium

Received April 2, 2003

The calculation of molecular quantum similarity measures using the molecular electron density requires the electron density and molecular alignment between two molecules. To obtain meaningful quantum similarity matrices, the electron density should be calculated efficiently and accurately and the alignment should be internally consistent. The internal consistency of the alignment for a series of molecules is investigated through distance geometry concepts. The calculation of the quantum similarity matrix requires the calculation of a quadratic number of similarity integrals, and a scheme to obtain these efficiently is developed. Both the alignment procedure and the ASA method for approximate molecular electron densities are tested for a set of steroid molecules.

INTRODUCTION

Molecular similarity and the way to express it in a quantitative measure is an active field of research due to the large number of applications, e.g., in database mining, 3-D pharmacophore searching, and so on. As a result, many algorithms have been devised to express the degree of molecular similarity, based on geometrical parameters such as dihedral angles or other molecular properties. Such similarity measures are then often used to interpret observed similarities in molecular properties or as a basis for clustering of molecules.[1−3]

From the fundamental postulates of wave function based quantum chemistry, it is clear that the basic entity describing the entire molecular system is the wave function. In DFT the basic entity is the electron density, which is naturally easily obtained from the wave function as well and contains the whole information one can obtain from the system. In recognition of this fundamental postulate, in the field of molecular quantum similarity, the similarity between two molecules is expressed via the similarity of the spatially smeared out electron density of the molecules involved. As such, Quantum Molecular Similarity is a technique able to compare any set of submicroscopic systems. More precisely, a *molecular quantum similarity measure* (*QSM*) between two molecules A and B can be expressed, among other several choices, as an overlap between the two electron densities in the integral measure[4−9]

$$Z_{AB} = \int \rho_A(\mathbf{r}_1)\rho_B(\mathbf{r}_1)d\mathbf{r}_1 \qquad (1)$$

where $\rho_A(\mathbf{r}_1)$ and $\rho_B(\mathbf{r}_1)$ refer to the electron density of molecules A and B, respectively, at the point $\mathbf{r}_1$ in space. For a set of M molecules, all elements Z_{AB} form a symmetrical ($M\times M$) matrix $\mathbf{Z}$, called the *molecular quantum similarity matrix* (*MQSM*).

The key feature of quantum similarity lies in the use of the electron density of a molecule. It is well-known that the electron density can be obtained through quantum chemical calculations. However, these usually are quite time-consuming, and as a consequence an approximate model for electron densities would be quite interesting. In the most recent work on quantum similarity and quantum QSAR, use is made of Atomic Shell Approximation electron[10−12] densities. In the present work, a comparison will be made between QSM values obtained using ASA approximate electron densities and true, ab initio electron densities. One of the important steps in the evaluation of the molecular quantum similarity measure introduced in eq 1 consists of the alignment of molecules A and B. Depending the mutual position of A and B, the integral measure in eq 1 will differ. This alignment step forms the second bottleneck of the evaluation of all QSM between two molecular structures in a set of M molecules. The most used techniques are structural alignment algorithms. A fine example is the Topo-Geometrical Superposition Approach (TGSA) introduced by Gironés et al.[13] An alternative approach consists of QSSA,[14] where the QSM maximization is pursued. It has been shown by Bultinck et al.[14] that such an approach yields higher molecular similarity measures and is more coherent with the idea of molecular quantum similarity. Moreover, QSSA can be used for any set of molecules or quantum objects, as long as the density functions are available. Topogeometrical approaches require the presence of sufficiently similar structural elements to function. This constitutes a limiting case for topogeometrical approaches. In the following, arguments are shown that

* Corresponding author phone: +32/9/264.44.23; fax: +32/9/264.49.83; e-mail: Patrick.Bultinck@UGent.be.
[†] Ghent University.
[‡] University of Girona.
[§] Universiteit Antwerpen.

10.1021/ci034060a CCC: $25.00
Published on Web 06/28/2003

structural alignment is not always consistent with QSM matrices, whereas QSSA alignment is. This approach will use other molecules as a reference to judge the consistency of the similarity index between two molecules. As such, the present consistency arguments are only applicable for sets of more than two molecules.

THEORETICAL BACKGROUND

The QSM between two molecules A and B, forming part of a set of M molecules, is expressed through the overlap integral shown in eq 1. For example in Hartree−Fock theory, it is immediately clear that this will require the calculation of a large set of integrals over the basis functions in molecules A and B.[15] This is an especially troublesome aspect in the evaluation of QSM since the mutual position of molecules A and B in space will also determine the QSM.

The problem of the enormous number of integrals, involved in the calculation of the overlap, may be reduced quite strongly by using the so-called Atomic Shell Approximation (ASA).[10−12] In this approach, the electron density in a molecule A with M_A atoms is approximated through a sum of atomic contributions:

$$\rho_A(\mathbf{r}) = \sum_{a=1}^{M_A} Z_a \rho_a^{ASA}(\mathbf{r}) \tag{2}$$

Each $\rho_a^{ASA}(\mathbf{r})$ represents the unity normalized electron density of a single atom a with atomic number Z_a. Within ASA the atomic electron densities are expanded in terms of a basis set of Gaussian s-type orbitals (GTO):

$$\rho_a^{ASA}(\mathbf{r}) = \sum_{i_a}^{N_G} w_{i_a} |s_{i_a}(\mathbf{r})|^2 \tag{3}$$

The conceptual similarity between this formula and the electron density over a set of MOs[15] can be easily seen. N_G denotes the number of s-type GTOs used in the expansion. It must be stressed that as the employed GTOs are chosen to be spherically symmetric s-type orbitals, the obtained density functions are rotationally invariant. The expansion coefficients $\{w_{i_a}\}$ are restricted to be convex: positive definite and summing the unit. The expansion coefficients as well as the GTO exponents were optimized by reproducing the electron density for every atom based on Hartree−Fock calculations. Using the coefficients and exponents published previously,[10,11] the total molecular density is obtained through eq 2. The use of ASA strongly reduces the computational effort for a single calculation of the QSM and moreover for repeated calculation of QSM as is done in the QSSA algorithm. The use of approximate electron densities may, however, influence the results for the QSM between two molecules and may even reverse similarity orders between different molecular combinations. It is therefore important to establish to what extent the ASA derived QSM differ from true ab initio QSM, where the ab initio electron density is used.

As described above, the value for the QSM also depends on the mutual position of both molecules in space. Boon et al.[16] have described a method where the alignment step is avoided, but that still uses quantum chemical ideas. Rather than using the smeared out electron density of the molecules involved, they used quantum chemically calculated electron density derived quantities such as Fukui functions and local softness. Using autocorrelation functions, the similarity is then calculated without need for any molecular alignment. This procedure is found to be handy but may give different results depending on the molecular descriptor used. Therefore, in the present study and in order to obtain stable results the electron density is used throughout but needing as a result molecular alignment.

Molecular alignment may be performed in several ways, structural or topogeometrical alignment being the most widely used. In the latter methods, molecules are aligned on the basis of structural features such as chemical bonds and sequences of chemical bonds. Most often a maximal number of atoms of the same element is superimposed in both molecules. When such an alignment is done, the QSM can be calculated readily from eq 1. One, however, immediately notices that different values for the QSM may be obtained, depending on the actual algorithm used for the structural alignment. A particularly often used method in quantum similarity measures is the TGSA method introduced by Gironés et al.[13] In this method molecules are aligned such that a maximum number of chemical bonds and sequences of two chemical bonds between the same elements in both molecules coincide. Since the QSM will depend on the actual implementation of the topogeometrical approach, there is some ambiguity in the resulting QSM.

Recently, Bultinck et al.[14] have introduced the QSSA approach, which abandons this geometrical path, and uses translational shifts and Euler angles to maximize directly the QSM. Such a procedure benefits from the coherent use of quantum similarity theoretical ideas both in molecular alignment as well as in computing similarity measures. It has been shown previously that this has some advantages over the more frequently used topogeometrical methods, consisting among others, in the fact that higher quantum similarity measures are often found when using QSSA and the possibility to align very dissimilar structures. QSSA alignment is also mathematically more consistent with the basic ideas of quantum similarity, because it attempts to obtain the unique upper limit of each element Z_{AB}, assuming nondegeneracy of the maximum. The QSSA method does, however, have the disadvantage of being a lengthier procedure, involving a genetic algorithm. As a consequence of the existence of many local maxima next to the global maximum for the QSM, there is the risk that when seeking for the global maximum, the optimization cannot get out of a local maximum. In the following, an algorithm is described to allow identifying such cases. More importantly, it also describes some fundamental requirements for alignment schemes to be used in quantum similarity. Even beyond this, these requirements should be tested for any case where molecular similarity, even not using quantum similarity, is used. This lies in the requirement of internal distance consistency, as described below.

RESULTS AND DISCUSSION

Molecular Similarity and Distance Measures. With all calculated QSM elements, the symmetric similarity matrix for a set of M molecules is constructed as

$$\mathbf{Z} = \left\{ Z_{IJ} \right\} \left(I, J = 1, M \right) = \begin{bmatrix} Z_{11} & Z_{12} & Z_{1M} \\ Z_{21} & Z_{22} & Z_{2M} \\ & & \\ Z_{M1} & Z_{M2} & Z_{MM} \end{bmatrix} \quad (4)$$

The diagonal elements Z_{II} are the so-called *quantum self-similarity measures*,[4−9] which naturally do not involve an alignment step. All the other nonredundant upper or lower triangle elements have to be obtained by calculation of the QSM, involving a specific method for molecular alignment. The main difference between a topogeometrical approach such as TGSA and QSSA lies in the fact that TGSA directly gives a certain alignment and the resulting QSM. In QSSA the alignment is produced by the maximization of the QSM. Beside the fact that QSSA gives higher QSM for the off-diagonal elements and that these QSM are unique upper values, apparently no other immediate advantages appear from the structure of **Z**.

In Euclidean mathematical space, use is often made of distance geometry concepts.[17] Clearly, this requires availability of a proper distance model. It is immediately clear that molecular QSM such as (1) are not to be confused as a distance but should be considered a scalar product, and as such distance geometry seems to be of no use in the maximization of the QSM. The molecular quantum similarity matrix (MQSM) **Z**, holding all elements Z_{AB} over the entire set of molecules can at first glance not be checked for internal consistency. Moreover, instead of the matrix **Z** itself, often a derived Carbó index matrix is used.[4−9] The Carbó index transforms the matrix **Z** in such a way, that all elements are in the interval]0,1] and reflect the extent of similarity between two molecules A and B:

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

Several years after Carbó et al.[18] studied the alternative definitions and the relationships between the distance and cosine-like Carbó indices, Maggiora et al. recently discussed the behavior of different similarity indices,[19] but most do not qualify as a true distance measure except under some well defined circumstances. Carbó et al. did, however, also introduce a direct quantum distance measure, namely the density difference squared module between two molecules A and B.[4] This involves the integral measure over entire space of the density difference:

$$d_{AB}^2 = \int (\rho_A(\mathbf{r}_1) - \rho_B(\mathbf{r}_1))^2 \, d\mathbf{r}_1 \quad (6)$$

It is immediately clear that the square root of the expression (6) corresponds to a Euclidean distance[17] defined in the Hilbert semispace of the density functions. Then, there exists an interesting relation between the distance matrix $\mathbf{D}=\{\mathrm{d_{IJ}}\}$ and the MQSM **Z**. Equation 6 leads to

$$d_{AB}^2 = \int \rho_A(\mathbf{r}_1)\rho_A(\mathbf{r}_1) d\mathbf{r}_1 + \int \rho_B(\mathbf{r}_1)\rho_B(\mathbf{r}_1) d\mathbf{r}_1 - 2 \int \rho_A(\mathbf{r}_1)\rho_B(\mathbf{r}_1) d\mathbf{r}_1 \quad (7)$$

$$d_{AB}^2 = Z_{AA} + Z_{BB} - 2Z_{AB} \quad (8)$$

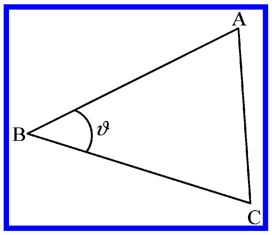


Figure 1. Triangle formed by three distances d_{AB}, d_{BC}, and d_{AC}.

If the essential features of the QSSA method are recalled, an attempt is made to maximize the similarity Z_{AB} by finding the optimal alignment. One immediately sees that the attempt to maximize the overlap quantum similarity measures Z_{AB} is equal to attempting the minimization of the squared distance d_{AB}^2. This is due to the fact that Z_{AA} and Z_{BB} do not involve a molecular alignment step, so they are independent of the alignment procedure used.

The use of a true squared distance derived from eq 8 does offer important advantages. Euclidean distances have the very interesting feature that they have to obey the distance inequalities as in classical geometry. In a triangle where the connecting lines are the distances d_{AB}, d_{AC}, and d_{BC}, as shown in Figure 1, the cosine rule holds:

$$d_{AC}^2 = d_{AB}^2 + d_{BC}^2 - 2d_{AB}d_{BC}\cos\vartheta \quad (9)$$

Given the fact that cosines are confined to the interval [-1,1], one immediately finds the well-known expressions for the lower and upper bounds for an unknown distance d_{AC} in terms of known values for d_{AB} and d_{BC}:

$$|d_{AB} - d_{BC}| \le d_{AC} \le d_{AB} + d_{BC} \quad (10)$$

$$d_{AC;B}^L \le d_{AC} \le d_{AC;B}^U \quad (11)$$

For every molecule B, also an element of the set of M molecules, one can calculate an upper bound $d_{AC;B}^U$ to the distance between A and C and a lower bound, denoted $d_{AC;B}^L$. Depending on the molecule B involved, a different value for the upper and the lower bound for the combination AC will be found. All distance geometry limits should naturally always be obeyed in all applications of distances between molecules, here expressed in the context of QSM.

In QSSA, one attempts the maximization of Z_{AB}. Assuming a single global maximum for the similarity, one will always have that the QSSA produced alignment will have a value that is lower than or equal to this maximum. Referring to eq 8 this means that the distance d_{AB} is always equal to or larger than the correct value. The correct value is that one produced if Z_{AB} is equal to the maximum value. This means that the distance between molecules A and B can easily be overestimated but never underestimated. Distances can easily be overestimated by failing to obtain the global maximum of the corresponding element of **Z**. This can efficiently be used as a tool to check the internal consistency of the alignments performed. For a specific combination of molecules A and C, eq 10 has to be obeyed for all B (other than A and C):

$$\forall\, B \neq \{A,C\}: d_{AC;B}^L \le d_{AC} \le d_{AC;B}^U \quad (12)$$

The most strict requirement over all B is the one with the lowest value for $d_{AC;B}^U$. If d_{AC} obeys this most strict requirement, then it automatically obeys all other requirements. Using the smallest value of $d^U_{AC} = \min_B \{d^U_{AC;B}\}$ over all B, one can build a matrix $\mathbf{D}^U = \{d^U_{AC}\}$. The value for the upper limit, d^U_{AC}, can then be related to a lower limit for the similarity between A and C via eq 8. This value, denoted Z^L_{AC}, provides an excellent way to check whether the alignment, produced by a combination A,C of molecules, is consistent with all the other alignments found using the same alignment procedure. Z^L_{AC} is the lower limit for the QSM for molecules A and C, and so the true value for Z_{AC} should be equal to or larger than this value. This means one can check the internal consistency of the MQSM by constructing the distance matrices **D** and the associated $\mathbf{D}^U$, deriving a matrix $\mathbf{Z}^L$ and checking whether the MQSM elements Z_{AC} are greater than or equal to Z^L_{AC}.

As was detailed before,[14] QSSA uses a Lamarckian genetic algorithm to align molecules to obtain a maximal QSM. Due to the existence of many local maxima during the optimization, and the fact that a genetic algorithm cannot ensure that the global maximum is found, it cannot easily be established whether one has obtained a sufficiently high level of confidence that the MQSM is chemically fit. However, using QSSA one can after the construction of the MQSM **Z** also construct the matrix $\mathbf{D}^U$. As a result, an efficient way to know whether the genetic algorithm is effectively producing a consistent result is obtained. The construction of the MQSM **Z** is repeated as long as the distance check flags any inconsistent elements. A newly found element of **Z** may also render some previously found value for a different element inconsistent, indicating a case where the genetic algorithm did not succeed in finding a sufficiently good alignment. This means that elements, which were at first identified as consistent, may be rendered inconsistent again when a new matrix $\mathbf{D}^U$ becomes available. In practice, this means that one calculates all elements of **Z** using QSSA and once these values are available, one can calculate a set of lower boundaries for all elements of **Z** and identify improper elements. As a result after every calculation of a matrix **Z**, the matrix $\mathbf{D}^U$ is constructed, and any elements of **Z** that are not consistent, flagged for further optimization.

The question may be raised whether it is possible in QSSA that internal distance consistency arises, even if the global maxima were not found. In principle this would be possible but rather unlikely. The genetic algorithm is not halted when a consistent alignment and QSM is found but rather is always carried out during a predefined number of generations. If after this number of generations, a consistent result is found, the genetic algorithm is stopped. If this is not the case, extra generations are produced until a consistent result is found. In most cases, even before the number of preset generations has been produced, a QSM value is found that is higher than the lower limit predicted by the distance inequalities. It is found that this new value may very well indicate previous alignments that are no longer consistent with this new finding. The genetic algorithm is then reinitiated for this previous alignment, until again this element is consistent with the boundaries. This procedure is repeated until the resulting matrix **Z** is consistent. Experience shows that in the consecutive steps the lower boundaries always increase, so that the difference between the value of an element Z_{AB} and the global maximum value for that element always decreases. Although this method to require internal consistency cannot guarantee that the global maximum will be found for each combination of molecules, it is observed that the method pushes the alignments to continuously produce higher QSM values, which obviously will reduce the difference with the global maximum. When by chance a couple of elements of **Z** with very high QSM are found, these yield stricter upper limits in eq 11, thereby pushing the other elements to higher values as well.

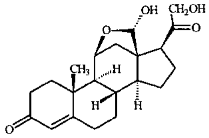
One can also address the lower limits in eq 11, thereby possibly opening the way to a matrix $\mathbf{D}^L$. This is much less straightforward due to the presence of the absolute value, the minus sign and the fact that none, any or both values of d_{AB} and d_{BC} may be overestimated in eq 10. Naturally, an upper limit for any element Z_{AB} is $\sqrt{Z_{AA}Z_{BB}}$.

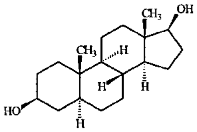
Although illustrated here mainly with the use for quantum similarity in mind, the above-described requirements should be met in all cases where molecular similarity is studied. If e.g. one expresses similarity as the agreement between calculated molecular descriptors and one can relate this to an Euclidean distance, this distance must obey the above-described distance inequalities. If not, there is a clear inconsistency in one of the parameters, and one needs to improve them or the way they were obtained. In the specific case of QSM such an inconsistency means that a specific element of **Z** is incorrect, which in turn means that one needs to search for a better alignment. The use of internally inconsistent matrices in any application should naturally be considered improper and be looked upon as preliminary steps subject to further improvement.

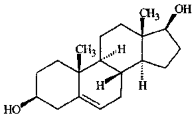
The implementation of the present distance inequalities is as follows. First all self-similarities are calculated for all molecules in the set. Then the MQSM **Z** is built element by element using the QSSA algorithm with a chosen minimal number of generations in the genetic algorithm used in QSSA. After the MQSM **Z** has been obtained, a matrix $\mathbf{D}^U$ is constructed. From this matrix the matrix $\mathbf{Z}^L$ is derived. If the actual value for Z_{AB} is below Z^L_{AB}, then the QSSA algorithm is reinitiated until a consistent matrix is found. Using the new values for the previously inconsistent elements, the new matrix **Z** is used to again obtain a new $\mathbf{Z}^L$ matrix, and the entire loop repeated. These cycles continue until all elements found are internally consistent. In case some previously obtained element is no longer consistent with the new matrix $\mathbf{Z}^L$, the QSSA algorithm is reinitiated to find a better alignment. In practice, it is found that during the search for internal distance geometry consistency, the elements of $\mathbf{D}^U$ continuously decrease, indicating that better alignments must be found before one can conclude consistency.

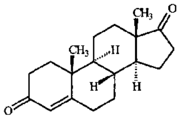
Application and Performance of ASA. As an example of the use of the agreement between ASA and ab initio calculated similarities and the distance considerations, a well-known set of steroid molecules was used to calculate the quantum similarity matrix **Z**. The actual set of molecules was used previously in several studies on QSAR and molecular similarity.[20-24] Lewis structures may be found in Table 1. The calculation of the QSM values for every combination of molecules involves the alignment of the molecular structures, and the calculation of the QSM using

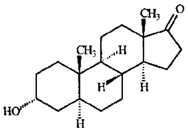
Table 1. Molecular Structures and Names for the 31 Molecules Contained in the Steroid Set

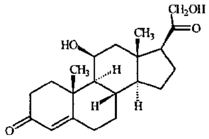

1
Aldosterone

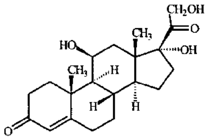

2
Androstanediol

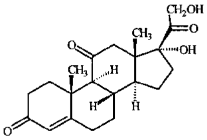

3
5-Androstanediol

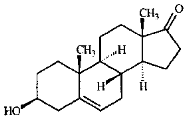

4
4-Androstenedione

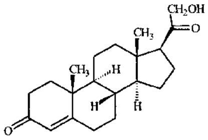

5
Androsterone

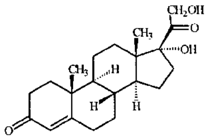

6
Corticosterone

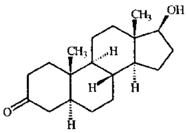

7
Cortisol

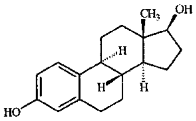

8
Cortisone

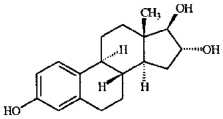

9
Dehydroepiandrosterone

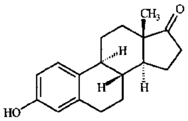

10
11-deoxycorticosterone

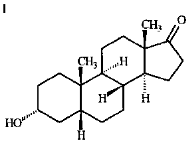

11
11-deoxycortisol

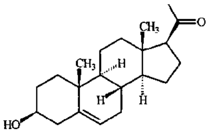

12
Dihydrotestosterone

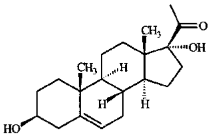

13
Estradiol

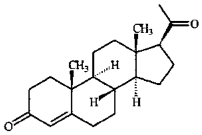

14
Estriol

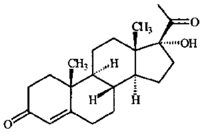

15
Estrone

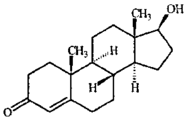

16
Ethiochonalonone

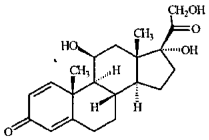

17
Pregnenolone

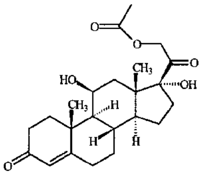

18
17a-Hydroxypregnenolone

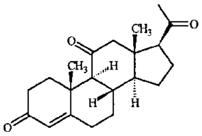

19
Progesterone

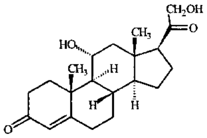

20
17a-Hydroxypregnenolone

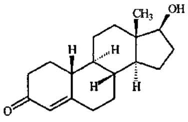

21
Testosterone

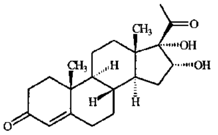

22
Prednisolone

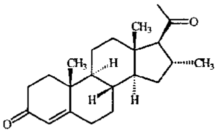

23
Cortisolacetat

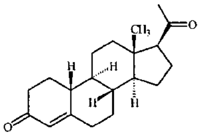

24
4-Pregnene-3,11,20-trione

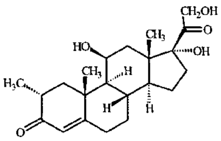

25
Epicorticosterone

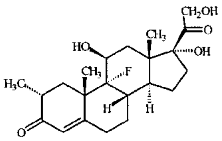

26
19-Nortestosterone


27
16a,17a-Dihydroxyprogesterone


28
17a-Methylprogesterone


29
19-Norprogesterone


30
2a-Methylcortisol


31
2a-Methyl-9a-Fluorocortisol

electron densities obtained through some electronic structure method.

First a comparison is made between the results obtained using a topogeometrical alignment method, TGSA and the previously described QSSA method for molecular alignment. For this application, electron densities are obtained using the ASA method. In second instance, the QSM obtained using ASA are compared to those obtained using ab initio calculations for the same molecular alignments.

Internal Consistency of MQSM in TGSA. In the TGSA method, which is a typical example of a topogeometrical approach, one tries to overlap as many structural elements as possible. These structural elements correspond to chemical bonds and sequences of two chemical bonds, always involving the same type of atoms in both molecules. Second, QSSA was used for molecular alignment by maximization of the QSM.

Using both techniques for molecular alignment, the MQSM $\mathbf{Z}$ was constructed, and an upper distance matrix $\mathbf{D}^U$ was constructed in the way described above. From this matrix $\mathbf{D}^U$, a lower bound matrix $\mathbf{Z}^L$ is constructed. The matrices $\mathbf{Z}$ and $\mathbf{Z}^L$ are then compared, and it is checked how many cases are present where the element of $\mathbf{Z}$ is not higher than that of $\mathbf{Z}^L$. Using the topogeometrical TGSA method, it was found that the matrix $\mathbf{D}^U$ is a very loose criterion. By this is meant that the upper limits in the distances are very high, and, as a result, the lower bounds in $\mathbf{Z}^L$ are so low that almost no violations are found, even it may be that $Z^L_{AB} = 0$, which is unphysical and impossible from the structure of Hilbert semispaces. At first glance this consistency may seem a good feature, but on the other hand it does leave a wide range for the similarity measures. That is, elements of $\mathbf{Z}$ may be quite strongly underestimated and still not give rise to violation of the distance boundaries. To illustrate this, consider the following example. Aldosterone and androstanediol (molecules 1 and 2 in Table 1) have a similarity of 287.8 when using TGSA. This value is consistent with the rest of the matrix. If this value is artificially lowered to 87.8, this is still consistent with the rest of the matrix. This manipulation does, however, induce an important chemical effect. Where originally the two molecules exhibited average similarity over all molecular combinations, namely a Carbó index of 0.31, it has dropped to less than 0.10 which indicates very poor similarity. This means that there is a very important degree of freedom in the QSM. When one manipulates the QSM for a certain combination AC as was done above, this can cause an artificial increase of the Euclidean distance as is clear from eq 8. But there are no molecules B which give a $d^U_{AC;B}$ that flags Z_{AC} as inconsistent. There is internal distance consistency because the limits for allowable Z_{AC} are so wide, especially the lower limit is much too low, even $Z^L_{AB} = 0$. This indicates that for the purpose of quantum similarity, great care should be taken in using topogeometrical alignments. $Z^L_{AB} = 0$ is impossible in any practical application and so indicates very unlikely chemical features. To illustrate different elements of the matrices introduced above, the reader is referred to the application of these matrices in the case of steroids (see below).

It is clear that every element Z_{AB} has an upper limit, namely corresponding to $Z_{AB} = \sqrt{Z_{AA}Z_{BB}}$. *This is a logical consequence of the fact that no quantum object can exhibit a higher similarity toward some reference quantum object than the reference quantum object to itself. In case of the optimal alignment a maximum value will be found which may lie closer to or further from the upper limit described above. QSSA actively seeks this maximum.* In TGSA a value for Z_{AB} is found, which usually can be rather far from this maximum. The derived distance d_{AB} is then overestimated when using TGSA compared to QSSA. The same goes for d_{BC}, so that $d^U_{AC;B}$ is too large. In fact, in TGSA it is found that these $d^U_{AC;B}$ are very often even larger than the largest possible value according to eq 8, namely

$$d_{AB} = \sqrt{(Z_{AA} + Z_{BB})} \tag{13}$$

where $Z_{AB} = 0$. Such a case is in practice impossible and serves only to obtain an upper limit for d_{AB}. When using TGSA, it is found that this limit is often stricter than any other limit induced by the $\{d^U_{AC;B}\}$.

Internal Consistency of MQSM in QSSA. When using the QSSA method in a first cycle for the different molecular combinations in the MQSM $\mathbf{Z}$, several inconsistencies were found. When using the boundaries to reinitiate the genetic algorithm, after a few cycles all these inconsistencies were removed, and the new alignments produced were found to yield quantum similarity measures above the threshold of the matrix $\mathbf{Z}^L$. It was also observed that when a new alignment is found for some element, some previous alignments and accompanying QSM were rendered inconsistent. In other words, a newly found value for an element of $\mathbf{Z}$ can cause the lower limit for a previously located element to grow larger than the element itself, causing a new inconsistency. This means that the entire procedure has to be repeated until consistency is reached for all elements. An important difference with TGSA is that during these repeated consistency cycles, one observes a continuing increase of the lower bounds for the QSM. The values obtained through QSSA for the QSM in $\mathbf{Z}$ are up to a factor 3 higher than when TGSA is used. More importantly, even minor modifications in the matrix $\mathbf{Z}$ obtained using QSSA are immediately flagged as inconsistent elements. There is much less flexibility in the QSM, and as such much less flexibility in the derived Carbó indices which give a better view of the degree of similarity. This is a key difference with TGSA. In QSSA internal consistency is found, not due to the fact that many distances are overestimated, but rather because the distance inequalities force the algorithm to continuously seek higher similarity alignments. A fine example lies in the following experiment. The TGSA alignments are taken as a starting point. These are internally consistent because all distances are strongly overestimated. Then two random alignments AB and BC are picked, and QSSA explores the hyperspace formed by the six alignment parameters to find higher similarity values. When this has been achieved for both combinations, the consistency of the element AC may be lost because $d^U_{AC;B}$ has highly decreased, and so Z^L_{AC} is no longer zero but adopts a higher value. So one can summarize that TGSA gives consistent results because the distances are overestimated, thereby creating a much too large interval for allowable values for Z_{AC}. On the other hand, QSSA in the first cycles gives inconsistent results because some values

Table 2: Example Applications of Distance Constraints for Different Molecular Combinations in the Steroid Set[a]

A	C	B	$Z_{AC}{}^{L}$	Z_{AC}
3	18	17	605.7	723.2
		17	303.5	519.4
3	4	9	214.7	488.3
		none	none	439.4
14	15	13	538.8	657.2
		13	378.7	530.4
22	26	25	466.1	466.1
		25	338.9	345.4
22	31	30	569.8	826.5
		7	224.1	535.3

[a] See text for explanation of symbols. Plain numbers refer to QSSA results, numbers in italics to TGSA results. “None” means that no B could be found that gives a physically meaningful Z^{L}_{AC}.

Z_{AC} do not fit with Z^{L}_{AC}. Consistency is then regained after higher QSM alignments are found, so that again Z_{AC} is above Z^{L}_{AC}. When the resulting matrix **D** does not exhibit any inconsistencies anymore, the matrix **Z** can, without problem, be used for subsequent quantum similarity based work, provided that the consistency is not an artifact due to overestimated distances. In the context of quantum QSAR (QQSAR), it should be mentioned that not the elements $\{Z_{IJ}\}$ are individually used as molecular descriptors but rather the entire columns of the matrix **Z** are used. The column Z_A of the matrix is used as molecular descriptor for molecule A within the set of M molecules. A particularly useful construct are stochastic transforms of the matrix **Z**.[25] In this case it should be argued that then the molecular descriptors are easily seen to no longer obey the inequalities of Euclidean space, but rather the metric is a Minkowski one.[17,26] In the latter space, the Euclidean triangular inequalities are no longer valid.[17] But since the derivation of the molecular descriptors involves a step in Euclidean space, the QSM should obey Euclidean inequalities.

As an illustration of the effect of distance geometry concepts, a number of examples from the steroid set are shown in Table 2. This table gives five combinations of two molecules (A and C in eq 12). For each combination, the molecule B which in eq 12 induces the strictest upper distance constraint is given, together with the induced value for Z^{L}_{AC}. Finally, the value found for Z_{AC} is given. These distance constraints may be derived both for **Z** matrices obtained using TGSA and QSSA, and both are given in Table 2.

Table 2 shows that in QSSA there are cases, where the fact that once a few high similarities are found, quite strict Z^{L}_{AC} values may be induced. In fact, several cases have been found where the TGSA alignment produces quantum similarities that do even not exceed the QSSA based lower bounds Z^{L}_{AC}. Also, one does not always find that the same B induces the strictest Z^{L}_{AC} for both alignment techniques. This is illustrated by combination 22−31. With QSSA reference molecule 30 induces a quite high lower bound for the similarity, whereas with TGSA, molecule 7 induces a quite loose lower bound. Another series of examples lies in those cases where, when using TGSA, no B is found that puts a physically meaningful limit on Z^{L}_{AC}. This is illustrated in Table 2 by the similarity between molecules 3 and 4.

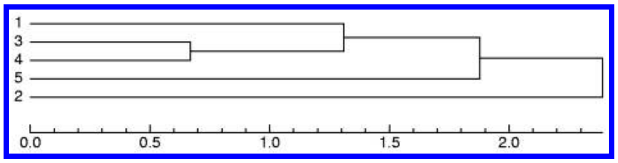


Figure 2. Dendrogram for a 5 × 5 steroid submatrix of the MQSM **Z** using TGSA based QSM.

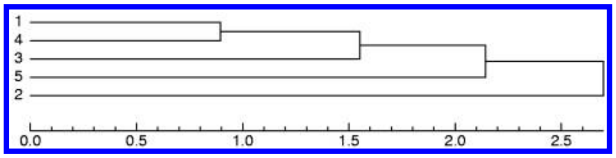


Figure 3. Dendrogram for a 5 × 5 steroid submatrix of the MQSM **Z** using QSSA based QSM.

It is worth mentioning that in all cases, when some high value for Z^{L}_{AC} is found, the QSSA genetic algorithm always finds an alignment so that $Z_{AC} \geq Z^{L}_{AC}$. No violations have been found against this, when using QSSA. When using TGSA, as commented before, no violations are found due to the fact that TGSA based similarities are quite often relatively far from the maximum similarity value, and as a result fairly loose distance constraints result. The similarity between molecules 14 and 15 is a nice example of this behavior. Using QSSA, one finds that the strictly minimal Carbó similarity index between molecules 14 and 15 would be 68.8%. With TGSA this is only 48.4%. This illustrates that TGSA is less strict. The QSSA algorithm as presented here, using the consistency checking to drive the genetic algorithm, ultimately reveals an alignment with a Carbó similarity index of 83.9%. It is worth noting too that the TGSA alignment produces a Carbó similarity index of 67.7%, which in fact would be rejected by QSSA since it does not fulfill the QSSA lower bound in the Carbó similarity index.

Influence of Molecular Alignment on Carbó Indices. An illustrative way to express the degree of similarity is through the Carbó index. This index is calculated as in eq 5 and has a value of 1 for perfect similarity with decreasing similarity indicated by lower values. The question may be raised if QSSA under the distance constraints gives a significantly different matrix of Carbó indices. It was found that when using QSSA approximately 39% of all elements of the transformed matrix is higher compared to the values in the TGSA derived matrix. Many of these elements do not differ to great extent, but there several cases where the Carbó index grows relevantly larger. As an example, the similarity between 5-androstanediol and pregnenolone (molecules 3 and 17 in Table 1) is rather strongly underestimated in TGSA, and a similar case exists for 5-androstanediol versus 17a-hydroxypregnenolone. The Carbó indices increase from 0.64 and 0.61 in TGSA to 0.89 and 0.85, respectively. This again indicates the importance of a good alignment algorithm where use is made of the distance constraints to continuously increase the similarity measures until consistency is achieved. To illustrate the influence of the alignment algorithm in a graphical way, Figures 2 and 3 give the dendrograms for a 5 × 5 submatrix of the MQSM **Z**. The molecules included are molecules 3, 5, 6, 17, and 24 from Table 1. For some combinations TGSA and QSSA give similar results, whereas for other combinations QSSA will give appreciably better results. The 5 × 5 Carbó transformed matrices **C** from TGSA

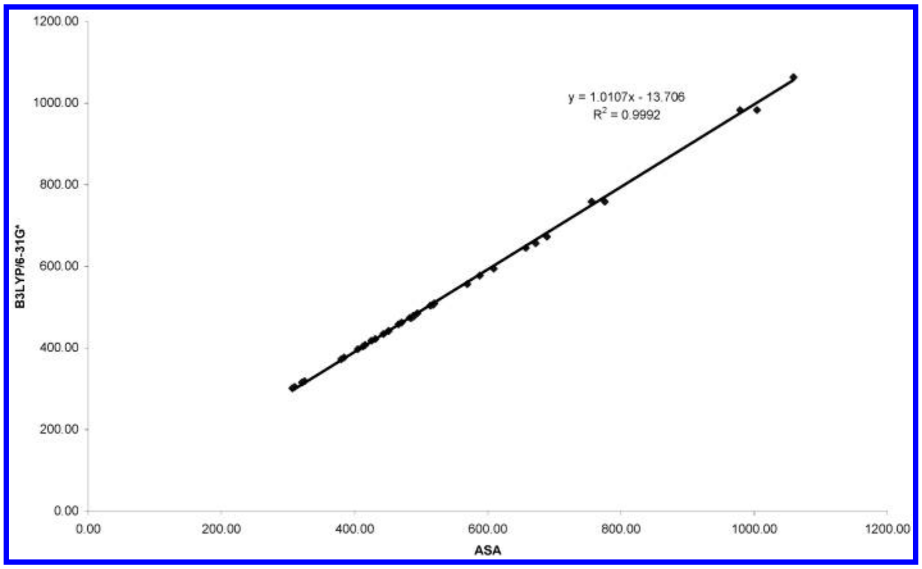


Figure 4. Agreement between B3LYP/6-31G* and ASA derived quantum similarity measures for molecules 1−9 of the steroid set.

and QSSA are also presented, and dendrograms were obtained from these Carbó similarities as described by Bultinck et al.[24]

$$\mathbf{C}_{TGSA} = \begin{bmatrix} 1.0000 & 0.4421 & 0.5320 & 0.6435 & 0.4221 \\ & 1.0000 & 0.4583 & 0.3954 & 0.3785 \\ & & 1.0000 & 0.6691 & 0.4957 \\ & & & 1.0000 & 0.5476 \\ & & & & 1.0000 \end{bmatrix} \quad (14)$$

$$\mathbf{C}_{QSSA} = \begin{bmatrix} 1.0000 & 0.4952 & 0.5868 & 0.8954 & 0.5238 \\ & 1.0000 & 0.4708 & 0.4747 & 0.4491 \\ & & 1.0000 & 0.6878 & 0.4957 \\ & & & 1.0000 & 0.5880 \\ & & & & 1.0000 \end{bmatrix} \quad (15)$$

From the figures and the matrices (14) and (15) it is clear that the influence of the alignment procedure is not negligible. Some elements of the Carbó matrices may not change to large extent, whereas other elements can undergo much bigger changes. Although for illustrative purposes limited to a 5 × 5 matrix which is still easily inspected visually, this effect is even more important if one considers the larger 31 × 31 matrices.

Applicability of ASA for QSM Calculations. When considering the values of the QSM, one can, beside the alignment algorithm, also investigate the performance of the ASA densities compared to true ab initio electron densities. One should not expect a completely perfect agreement between the two QSM derived from two types of electron density, but the agreement should be sufficiently good not to reverse molecules in similarity ordering. For steroid molecules 1−9, B3LYP/6-31G* electron densities were calculated using the Gaussian-98 program.[27] This level of calculation is generally accepted as a sufficiently high level of calculation in computational medicinal chemistry. The ab initio QSM were then calculated for the 9 × 9 similarity matrix using the BRABO ab initio program.[28] The agreement between the B3LYP/6-31G* and ASA derived similarity measures is depicted graphically in Figure 4.

As Figure 4 clearly shows, the agreement is excellent. The correlation coefficient is above 99.9%, and the ASA approximate electron densities give very good quantum similarities. The gain in computer time is enormous when using ASA versus B3LYP/6-31G*, since one does not need to calculate ab initio electron density overlap. ASA similarities can be calculated in a matter of milliseconds on a current Linux based Pentium IV PC. The calculation of the overlap similarity using B3LYP/6-31G* densities requires first the ab initio calculations which is a very time-consuming step.

The question could be raised whether the excellent agreement of Figure 4 is somehow fortuitous. As has been shown previously, ASA electron densities are quite accurate for large molecules.[12] Further proof of the quality of the ASA method can be found in the following application. Considering eq 2, one sees that the electron density is a sum of atomic contributions. For larger molecules such as steroids, where the number of atoms also differs between different molecules, one can expect that for a set of different molecules, there will be a fairly large range of values for the elements Z_{AB}. This is indeed found in Figure 4. A critical test exists in checking whether a good correlation between ab initio and ASA similarities is still found when the range of values is much smaller. A fine example of a set of molecules where such behavior may be expected lies in a set of bond isomers. Bultinck et al.[14] have previously reported the similarity matrix for a noncongener set of seven bond isomers with the stoichiometry $C_4H_6O_2$. Structures of the molecules involved may be found in Bultinck et al.[14] The $\mathbf{Z}_{QSSA}$ matrix clearly shows that expressing the electron densities as a sum of atomic densities for binding isomers results in a much smaller range of elements Z_{AB}. This is especially clear if one considers the diagonal elements of the similarity matrix:

$$\mathbf{Z}_{QSSA} = \begin{bmatrix} 293 & 145 & 129 & 224 & 222 & 226 & 117 \\ & 293 & 169 & 185 & 137 & 149 & 115 \\ & & 293 & 160 & 170 & 134 & 148 \\ & & & 293 & 213 & 228 & 119 \\ & & & & 293 & 217 & 131 \\ & & & & & 293 & 121 \\ & & & & & & 293 \end{bmatrix} \quad (16)$$

For the different elements of $\mathbf{Z}_{QSSA}$ values were also calculated using B3LYP/6-31G* electron densities. The

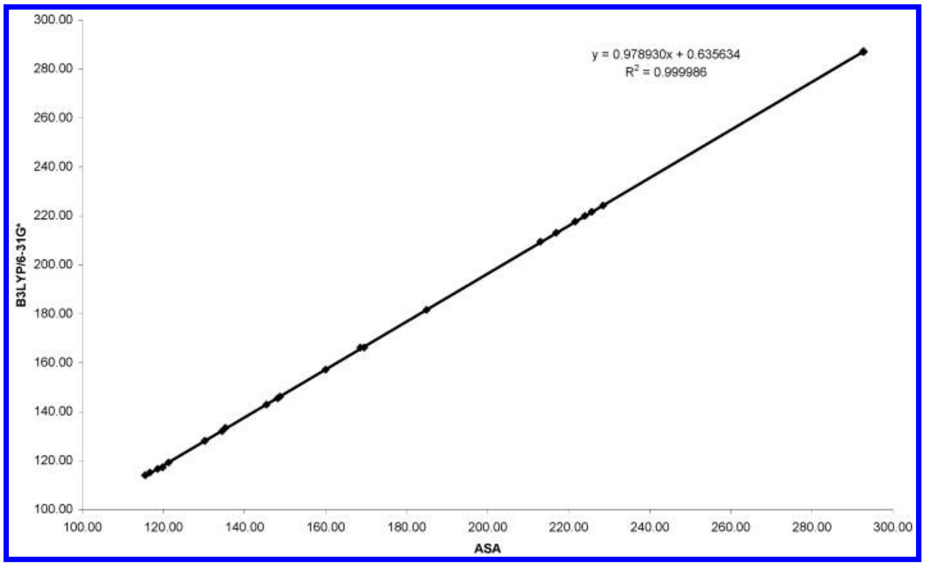


Figure 5. Agreement between B3LYP/6-31G* and ASA derived quantum similarity measures for $C_4H_6O_2$ binding isomers.

agreement between both sets of values is shown in Figure 5.

This figure again provides clear justification for the use of ASA approximate electron densities for the calculation of both self-similarities as QSM between two different molecules, even for cases where high accuracy is required to discriminate between nearly equal values.

CONCLUSIONS

Using molecular quantum similarity measures as an overlap integral of electron densities between two molecules, electron density Euclidean distances can be constructed. These are clearly defined as Euclidean norms, and as such should obey distance geometry relations.

An algorithm was established, which allows obtaining upper limits for the distance between the two molecules involved in the QSM. Using the relation between this electron density Euclidean distance and molecular quantum similarity measures, one can calculate a lower limit for the molecular quantum similarity between the two molecules involved. This allows identifying instances where calculated quantum similarity measures are insufficiently maximized, indicating that the globally optimal alignment was not yet located.

Using a test set of 31 steroid molecules, the construction of a lower bounds matrix for the quantum similarity matrix elements was implemented as part of the QSSA program. This approach was found to yield a complete internally consistent molecular quantum similarity matrix. In comparison to quantum similarity matrices constructed using topo-geometrical molecular alignments, the present algorithm induces much higher values for the QSM, yielding in this way nearest optimal QSM.

Finally, it is argued that for all applications of molecular similarity the internal consistency should be checked. If a consistent similarity matrix is not found, this indicates that one or more aspects of the similarity calculation exhibit flaws. One should then look for the reliability of these parameters, and aim to improve them. In the present quantum similarity case, the parameters are the elements of the quantum similarity matrix, and the flaw has to be found in an insufficiently good molecular alignment.

Overlap quantum similarity measures require the availability of molecular electron densities. For the set of steroid molecules a comparison was made between true ab initio electron density derived QSM and those obtained using ASA approximate densities. The results show that the agreement is excellent.

ACKNOWLEDGMENT

P. Bultinck wishes to thank the *Fund for Scientific Research-Flanders* (Belgium) for their grants to the Ghent Quantum Chemistry Group and acknowledges the *European Community − Access to Research Infrastructure action* of the *Improving Human Potential Program* for 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.

Supporting Information Available: The $\mathbf{Z}_{TGSA}$ and $\mathbf{Z}_{QSSA}$ matrices together with the derived matrices as obtained using TGSA and QSSA for the set of steroid molecules. This material is available free of charge via the Internet at http://pubs.acs.org.

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 2 molecular-structures. *Int. J. Quantum Chem.* **1980**, *17*, 1185−1189.

(5) Carbó-Dorca, R.; Besalú, E. A general survey of Molecular Quantum Similarity. *J. Mol. Struct. (THEOCHEM)* **1998**, *451*, 11−23.

(6) Carbó-Dorca, R.; Robert, D.; Amat, L.; Gironés, X.; Besalú, E. Molecular Quantum Similarity in QSAR and Drug Design. *Lecture Notes Chem.* **2000**, *73*.

(7) 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.

(8) 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.

(9) Besalu, E.; Girones, X.; Amat, L.; Carbó-Dorca, R. Molecular quantum similarity and the fundamentals of QSAR. *Acc. Chem. Res.* **2002**, *35*, 289−295.

(10) Amat, L.; Carbó-Dorca, R. Use of promolecular ASA density functions as a general algorithm to obtain starting MO in SCF calculations. *Int. J. Quantum Chem.* **2002**, *87*, 59−67.

(11) 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.

(12) Girones, X.; Amat, L.; Carbó-Dorca, R. Modeling large macromolecular structures using promolecular densities. **2002**, *42*, 847−852.

(13) Girones, X.; Robert, D.; Carbó-Dorca, R. TGSA: A molecular superposition program based on topo-geometrical considerations. *J. Comput. Chem.* **2001**, *22*, 255−263.

(14) Bultinck, P.; Kuppens, T.; Gironés, X.; Carbó-Dorca, R. Quantum similarity superpostion algorithm (QSSA): a consistent scheme for molecular alignment and molecular similarity based on quantum chemistry. *J. Chem. Inf. Comput. Sci.*, **2003**, accepted for publication.

(15) *Modern Quantum Chemistry*; Szabo, A.; Ostlund, N. S.; Dover Publications: New York, 1996.

(16) Boon, G.; Langenaeker, W.; De Proft, F.; De Winter, H.; Tollenaere, J. P.; Geerlings, P. Systematic study of the quality of various quantum similarity descriptors. Use of the autocorrelation function and principal components analysis. *J. Phys. Chem. A* **2001**, 8805−8814.

(17) *The CRC Concise Encyclopedia of Mathematics*; Weisstein, E. W., Ed.; CRC Press: 1998.

(18) Carbó, R.; Besalú, E.; Amat, Ll.; Fradera, X. On Quantum Molecular Similarity Measures (QMSM) and Indices (QMSI). *J. Math. Chem.* **1996**, *19*, 47−56.

(19) Maggiora, G. M.; Petke, J. D.; Mestres, J. A general analysis of field-based molecular similarity indices. *J. Math. Chem.* **2002**, *31*, 251−270.

(20) 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.

(21) 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.

(22) 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.

(23) 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.

(24) Bultinck, P.; Carbó-Dorca, R. Molecular quantum similarity matrix based clustering of molecules using dendrograms. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 170−177.

(25) 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.

(26) Carbó-Dorca, R. Shell partition and metric semispaces: Minkowski norms, root scalar products, distances and cosines of arbitrary order. *J. Math. Chem.* **2002**, *32*, 201−223.

(27) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery Jr., J. A.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; Kudin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Baboul, A. G.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Gonzalez, C.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Andres, J. L.; Gonzalez, C.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A. *Gaussian 98*, Revision A.7; Gaussian, Inc.: Pittsburgh, PA, 1998.

(28) Van Alsenoy, C.; Peeters, A. Brabo − a program for ab initio studies on large molecular-systems. *J. Mol. Struct. (THEOCHEM)* **1993**, *286*, 19−34.

CI034060A