Quantum Similarity Superposition Algorithm (QSSA): A Consistent Scheme for Molecular Alignment and Molecular Similarity Based on Quantum Chemistry

Patrick Bultinck,*,† Tom Kuppens,† Xavier Gironés,‡ and Ramon Carbó-Dorca‡

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 January 27, 2003

The use of the molecular quantum similarity overlap measure for molecular alignment is investigated. A new algorithm is presented, the quantum similarity superposition algorithm (QSSA), expressing the relative positions of two molecules in terms of mutual translation in three Cartesian directions and three Euler angles. The quantum similarity overlap is then used to optimize the mutual positions of the molecules. A comparison is made with TGSA, a topogeometrical approach, and the influence of differences on molecular clustering is discussed.

INTRODUCTION

Concepts such as molecular similarity are used throughout chemistry. When two molecules react in very much the same way in some reaction, organic chemists often try to relate this to the "similar" structure or properties of the two molecules or parts thereof, e.g., the similarity in leaving group and its surroundings in a nucleophilic reaction. Although used all over chemistry, despite its simplicity at first glance, a concept like molecular similarity is not so easily described in a mathematical form based on quantum chemical ideas.

Very much the same goes for the alignment of molecules or the alignment of a smaller molecule within a larger molecule. Again, in a logical visual approach one tries to find spaces within the two molecules where a high similarity becomes apparent. From the above, it is immediately seen that both concepts are strongly related to each other. Many schemes have already been derived to quantify the similarity of two molecules and to align molecules.[1−3]

Most alignment procedures in common use superpose molecules on the basis of their molecular geometries. Such topogeometrical approaches usually involve the identification of sets of corresponding atoms a and b in molecules A and B, respectively. Then, topogeometrical approaches usually try to align the molecules in such a way that the maximum number of such atomic couples coincide. An interesting approach, often used in the field of quantum QSAR, is the topogeometrical TGSA method developed by Gironés et al.[4] There, an efficient structural alignment procedure is used. Based on the similarity between atomic dyads and triads in both molecules, it manages to align both structures in a convenient way. However, the most limiting feature of TGSA and many other topogeometrical approaches consists of the fact that the different molecules must somehow form a congener set or at least show some degree of topographical similarity. When such similarity is absent, then this constitutes a limiting problem, and as a consequence TGSA cannot be applied in this case.

The aim of the present work is to derive a general and internally consistent scheme for the calculation of molecular similarity and alignment, based on quantum chemical ideas, which will take into account the above-described link between the two concepts. Such a scheme should also be able to treat molecules which do not exhibit large topographical similarity, thereby allowing the alignment and similarity calculation for noncongener molecules.

In wave function based quantum chemistry the all-determining entity is the wave function itself. Nevertheless, this mathematical object does not carry any physical meaning. In density functional theory, per force, the all determining element is the electron density, which does have physical meaning. The electron density is also readily obtained from the wave function, naturally, so in the present work we will build an approximate electron density in order to derive similarity and alignment equations and algorithms. Next to these theoretical considerations, the scheme will be tested on two sets of molecules, and differences with TGSA alignments discussed as well as the possible consequences for performing molecular clustering and QSAR calculations.

MOLECULAR SIMILARITY

In the following section only a brief review of the already published aspects of molecular quantum similarity (MQS) will be presented. MQS was defined for the first time in 1980.[5] More in depth reviews can be found in Carbó et al.[6−10] It suffices to note that the quantum similarity measure of two molecules A and B can be simply obtained through the integral measure

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

where $\rho_A(\mathbf{r}_1)$ refers to the electron density of molecule A at

* Corresponding author phone: +32/9/264.44.23; fax: +32/9/264.49.83; e-mail: Patrick.Bultinck@UGentbe.
† Ghent University.
‡ University of Girona.

10.1021/ci0340153 CCC: $25.00
Published on Web 05/13/2003

some point $\mathbf{r}_1$ in space. For a set of N molecules, all elements Z_{AB} form a symmetrical ($N\times N$) matrix $\mathbf{Z}$, called the *quantum similarity matrix*. The operator $\Omega(\mathbf{r}_1,\mathbf{r}_2)$ is always chosen as a positive definite operator or a convex combination of these. Examples are the kinetic energy operator[11] or the coulomb operator.[12] In the present work the Dirac delta function, $\delta(\mathbf{r}_1 - \mathbf{r}_2)$, is used, so that eq 1 becomes

$$Z_{AB} = \int \rho_A(\mathbf{r}_1)\rho_B(\mathbf{r}_1)d\mathbf{r}_1 \tag{2}$$

This is in fact an overlap integral between the electron densities of molecules A and B. These can be obtained through a quantum chemical calculation directly, but this usually involves a very high computational effort. A useful alternative is the so-called Promolecular Atomic Shell Approximation (PASA).[13] The electron density of a molecule with M_A atoms is approximated as the sum of the electron densities of the atoms composing the molecule, that is

$$\rho_A(\mathbf{r}) = \sum_{a=1}^{M_A} Z_a \rho_a^{ASA}(\mathbf{r}) \tag{3}$$

Each $\rho_a^{ASA}(\mathbf{r})$ represents the unity normalized electron density of a single atom a, attached to an atomic number Z_a. Within the Atomic Shell Approximation (ASA)[14] the atomic electron densities are expanded in terms of a basis set of spherical Gaussian Type Orbitals (GTO):

$$\rho_a{}^{ASA}(\mathbf{r}) = \sum_{i_a}^{N_G} w_{i_a} |s_{i_a}(\mathbf{r})|^2 \tag{4}$$

Note the conceptual similarity between this formula and the electron density over a set of MOs.[15] N_G denotes the number of GTOs of *s* symmetry used in the expansion. It must be stressed that the employed GTOs are all chosen as spherically symmetric s-type orbitals, which yields rotational invariant density functions. The expansion coefficients $\{w_{i_a}\}$ are restricted to be convex: positive definite and summing the unit. Such coefficient conditioning is the most important feature of the ASA approach as will be discussed below. The ASA densities are also required to be normalized in the usual statistical sense:

$$\int \rho_a{}^{ASA}(\mathbf{r})d\mathbf{r} = 1 \tag{5}$$

From eqs 4 and 5 one can deduce that, provided the GTO are normalized, there must be present an additional requirement for the expansion coefficients, namely:

$$\sum_{i_a}^{N_G} w_{i_a} = 1 \tag{6}$$

Values for the expansion coefficients and exponents for the GTOs have been previously reported.[16] Both ASA and PASA techniques have been shown to be quite effective methods to obtain good quality electron densities for atoms and molecules respectively.[17] In the following, ASA GTO exponents and coefficients are taken from a fit to Hartree−Fock 6-311G electron densities.[16] The main advantage naturally consists of the fact that the usual bottlenecks of molecular integrals are avoided, and the similarity measure defined in eq 2 becomes a quite simple expression involving only s-type GTOs:

$$Z_{AB} = \int_{-\infty}^{\infty} \left(\sum_a^{M_A} Z_a \sum_{i_a}^{N_G} w_{i_a} |s_{i_a}(\mathbf{r})|^2\right)\left(\sum_b^{M_B} Z_b \sum_{i_b}^{N'_G} w_{i_b} |s_{i_b}(\mathbf{r})|^2\right) d\mathbf{r}$$

$$= \sum_a^{M_A}\sum_b^{M_B} Z_a Z_b \sum_{i_a}^{N_G}\sum_{i_b}^{N'_G} w_{i_a} w_{i_b} \int_{-\infty}^{\infty} |s_{i_a}(\mathbf{r})|^2 |s_{i_b}(\mathbf{r})|^2 d\mathbf{r} \tag{7}$$

Integrals over products of s-type GTOs are quite easily solved, since such GTO products yield another s-type GTO themselves, and thus, the integrals are transformed into easily solved overlap integrals over s-type functions.[15]

Although these derivations are quite simple, it is important to take into account that the results of eqs 2 and 7 will depend on the alignment of the involved pair of molecular structures. Depending on the relative orientation of both molecules, their electron densities will differ in the same point in space in an external coordinate system. This becomes an important problem in quantum similarity. The most immediate and at the same time mathematically correct approach would be to use, as an alignment criterion, the maximal value of the similarity measure Z_{AB}. Aligning the molecules then comes down to searching for the relative orientation of molecule B versus molecule A where Z_{AB} becomes maximal. Unfortunately, such an approach may represent a high computational effort. Constans et al. have presented an algorithm in which they approach atomic densities collapsed at the nuclei as a first step and apply a global search technique to find a superposition that maximizes Z_{AB}.[18] This is, however, a lengthy procedure, and the idea of collapsing the electron density in the nuclei is not physically sound from the quantum mechanical point of view.

MOLECULAR ALIGNMENT

Molecular Alignment. As is clear from the above discussion, no efficient and physically appealing method to superimpose molecules in a general way for any set of molecules has yet been described. In the following part, an algorithm is described, which uses the maximum quantum similarity measure as a conditioned way to align molecules. First, an algorithm should be well set, which explicitly expresses Z_{AB} in terms of some parameters describing the mutual molecular positions.

The relative positions of two molecules A and B in three-dimensional space can be described by a set of six parameters. The easiest to handle is the translation. Suppose molecule A is fixed in space, possessing some fixed conformation, then its center of mass also remains fixed. The second molecule, also in a fixed conformation, is then moved through space in order to maximize the molecular similarity Z_{AB}. With respect to molecule A, the center of mass of molecule B can only exhibit a translation. Such a translation can be decomposed in terms of *X*, *Y*, and *Z* components, relative to the A coordinate system. These components will be denoted as $\{T_X,T_Y,T_Z\}$. Next to the translation, one should also consider the rotational degrees of freedom of molecule B in the A coordinate system. Such a coordinate transformation is most easily described through the so-called Euler angles.[19] In what follows, the ZYZ order of Euler angles

will be used. This choice is by no means unique, and six possible sequences of rotations may be distinguished. The ZYZ order means that first a rotation around the Z axis of molecule B is performed over an angle α. Then a rotation around Y occurs over an angle β, and eventually a rotation occurs around Z over an angle γ. This means that the position (x_b^0,y_b^0,z_b^0) of an atom b of molecule B in the coordinate system of molecule B can be transformed to the coordinate system of molecule A by the following transformation:

$$\begin{bmatrix} x_b \\ y_b \\ z_b \end{bmatrix} = \begin{bmatrix} T_X \\ T_Y \\ T_Z \end{bmatrix} + \begin{bmatrix} \cos\alpha\cos\beta\cos\gamma - \sin\alpha\sin\gamma & \sin\alpha\cos\beta\cos\gamma + \cos\alpha\sin\gamma & -\sin\beta\cos\gamma \\ -\cos\alpha\cos\beta\sin\gamma - \sin\alpha\cos\gamma & -\sin\alpha\cos\beta\sin\gamma + \cos\alpha\cos\gamma & \sin\beta\sin\gamma \\ \cos\alpha\sin\beta & \sin\alpha\sin\beta & \cos\beta \end{bmatrix} \begin{bmatrix} x_b^0 \\ y_b^0 \\ z_b^0 \end{bmatrix} \quad (8)$$

It should be noted that the calculation of the PASA for molecule B is best carried out after the coordinate system transformation, since in this case the transformation only has to be carried out for the coordinates of the atoms. After this transformation, the PASA of molecule B can be calculated through an equation similar to (3).

Maximization of the Quantum Similarity Measure. One of the problems in molecular alignment and MQS maximization corresponds to the fact that there appear numerous local maxima in the similarity measure Z_{AB} with respect to the displacement coordinates $\{T_X,T_Y,T_Z\}$ and the Euler angles $\{\alpha,\beta,\gamma\}$. At first glance, one would expect a fairly simple optimization, since the dimensionality of the optimization space is low, as there are only six parameters involved. As will be described below, the optimization is not so straightforward, even with the use of a Lamarckian genetic algorithm. Prior to describing the actual algorithm, it is worthwhile to examine some features of molecular similarity and alignment.

Preliminary Considerations. The first simple question before proceeding to maximization can be associated to the domains in a molecule where the largest electron density is found. The highest electron density is well-known to reside in the subvalence or core region of the atoms. Intuitively, one expects that the integral Z_{AB} in eq 2 will increase when the core of an atom of molecule A coincides with the core of an atom in molecule B. Moreover, the quantum similarity measure integral will grow larger as more atoms of molecule A become coincident with atoms of molecule B. If the essential features of many geometric alignment methods, including the TGSA method,[4] are recalled, then one can realize that these topographical schemes follow a similar procedure. TGSA, using atomic dyads and triads, attempts to superimpose as many as possible atoms and chemical bonds. In those cases where an equal number of dyads and triads can be identified, in TGSA a score is calculated by taking the sum of the squared differences between the coordinates of all atoms in A and B:

$$d_{AB} = \sum_{\substack{a=1 \\ a\in A}}^{M_A} \sum_{\substack{b=1 \\ b\in B}}^{M_B} |r_a - r_b|^2 \quad (9)$$

The meaning of expression (9) consists of the fact that smaller values indicate better alignments. In the actual TGSA implementation d_{AB} is not used directly as a measure, but rather a derived Carbó-like index[20] is employed. For a set of congener molecules, the TGSA method is expected to yield quite high molecular similarity values because a maximal number of atomic cores will be aligned. A good TGSA alignment might then be expected to result in a fairly large quantum similarity measure.

The question that may be raised is where the advantage of the presently described method resides. Although one intuitively feels that, especially for congener molecules, maximum overlap similarity measures will be found when the atoms maximally coincide; there are some advantages, which can be attached to the present QSM algorithm. In TGSA one attempts to maximally superimpose nuclei of the same chemical element and bonds constituted of atoms of the same elements. As eq 7 shows, maximization of the QSM does not require alignment of atoms of the same element. This means that the maximization of QSM can be performed even if the involved molecules are not congener. One can thus in fact investigate how good TGSA performs, and starting from a TGSA optimal alignment, the quantum similarity measure Z_{AB} can be optimized, using the present technique. This comes down to investigating how large the TGSA error is by releasing the requirement of alignment of atoms of the same element and chemical bonds constituted of atoms of the same kind. But more importantly, one can also handle molecules where TGSA is *not* applicable, due to the absence of enough common structural features as dyads and triads. In this manner, one can hypothesize that when congener molecules are studied, TGSA can be used as a first methodological alignment step. After that, the finer technique presented here, can be employed. But even in the case of congener molecules, TGSA should not be applied without further checking within the present approach, as will be demonstrated below.

The QSSA Algorithm. The actual implementation of the Z_{AB} maximization method is based on a Lamarckian genetic algorithm. Such a technique has been used previously by Bultinck et al. for difficult optimizations with many local extrema.[21,22] First consider eq 7 again, writing out explicitly the integral:

$$Z_{AB} = \sum_a^{M_A}\sum_b^{M_B} Z_a Z_b \sum_{i_a}^{N_G}\sum_{i_b}^{N'_G} w_{i_a} w_{i_b} \int_{-\infty}^{\infty} |s_{i_a}(\mathbf{r})|^2 |s_{i_b}(\mathbf{r})|^2 d\mathbf{r} \quad (10)$$

Any GTO consists of a Gaussian function multiplied by its normalization constant. Using the well-known expression for the Gaussian normalization constant, eq 10 becomes

$$Z_{AB} = \sum_a^{M_A}\sum_b^{M_B} Z_a Z_b \sum_{i_a}^{N_G}\sum_{i_b}^{N'_G} w_{i_a} w_{i_b} \left(\frac{2\xi_{i_a}}{\pi}\right)^{3/2} \left(\frac{2\xi_{i_b}}{\pi}\right)^{3/2} \int_{-\infty}^{\infty} |e^{-\xi_{ia}\mathbf{r}^2}|^2 |e^{-\xi_{ib}\mathbf{r}^2}|^2 \, d\mathbf{r} \quad (11)$$

The latter integral is well-known, and eq 11 calling

$$\gamma_{i_a i_b} = 2\xi_{i_a}\xi_{i_b}(\xi_{i_a} + \xi_{i_b})^{-1}$$

is transformed into

$$Z_{AB} = \sum_{a}^{M_A}\sum_{b}^{M_B} Z_a Z_b \sum_{i_a}^{N_G}\sum_{i_b}^{N'_G} \left(\frac{2\xi_{i_a}}{\pi}\right)^{3/2} \left(\frac{2\xi_{i_b}}{\pi}\right)^{3/2} w_{i_a} w_{i_b} (\pi^{-1}\gamma_{i_a i_b})^{3/2} K_{i_a i_b} \quad (12)$$

where the following definition has to be taken into account:

$$K_{i_a i_b} = \exp(-\gamma_{i_a i_b} |R_a - R_b|^2) \quad (13)$$

Equations 12 and 13 are quite interesting for optimization purposes because they show that the only set of terms that depend on the actual molecular alignment are the functions $\{K_{i_a i_b}\}$ as given in eq 13. Thus, only these terms need to be optimized with respect to the translation parameters and Euler angles.

Since the maximization of the molecular quantum similarity measure involves optimizing six parameters, one could expect a fairly simple optimization algorithm. Experience shows that this is not always the case and that many local maxima exist. This is naturally similar to the general problem of conformational searching, where also one attempts to find the global minimum in a large collection of local minima.[23]

Starting from a certain random guess of the six alignment parameters the quantum similarity measure is locally optimized by using the simplex method. This first step may yield a local maximum similarity alignment, but nothing guarantees that the simplex will move to the global maximum. One could then opt to generate many random sets of alignment parameters and optimize all of these using the simplex method. Such an approach is not only computationally demanding, but the chances of finding the global maximum by such random approaches is quite small.

An interesting scheme to accelerate the global maximization of the molecular quantum similarity measure lies in the use of a genetic algorithm. An accessible introduction to genetic algorithms may be found in Leach.[23] In this method, consecutive generations have an increasing average fitness, which in the present case is related to an increasing average similarity over all members of the population. Also, usually the highest similarity over all members of the population increases as a result of the genetic algorithm procedure. In the "breeding" step, where a roulette wheel selection scheme is used, prior to proceeding to generating new offspring, virtual individuals are created using the well-known techniques as crossover, mutation, ... All virtual children are then locally optimized using the simplex method, and the resulting sets of parameters are then used as parents for the next breeding step. This method of implementing an intermediate local optimizer within a genetic algorithm is known as a Lamarckian genetic algorithm, which may speed up genetic algorithms quite drastically. In the current implementation, 100 randomly generated sets of alignment parameters are produced and optimized using the simplex method. These 100 optimized sets are then used as a first generation in the genetic algorithm. In every generation 10 new random sets of parameters are introduced, mimicking migration of new species in the population. In every breeding step, all members of the population are locally optimized using the simplex method, prior to using them as parents. The question may be raised why use is made of the simplex method, rather than a faster method. Such a method may consist in the calculation of the gradient and Hessian of Z_{AB} in terms of the alignment parameters. If these become available, one can use e.g., a Newton−Raphson scheme for the optimization of Z_{AB}. Since the electron densities of the molecules are approached through a linear combination of S-type GTOs, one can obtain with relative easiness explicit expressions for the gradient and Hessian matrix elements.[24] Unfortunately, computational practice shows this gradient-Hessian approach to be very problematic. This is in part due to the presence in the QSM gradient and Hessian equations of both a function increasing when atoms are more coincident, namely the exponential Gaussian function part as well as a decreasing function, namely the linear difference in Cartesian coordinates. Another problem corresponds to the coupling between alignment parameters, in the sense that e.g., the x_b coordinate in eq 8 is influenced by both T_X, α, β, and γ. This makes the Newton−Raphson scheme that was initially implemented highly troublesome. The simplex method, although demanding and slow in terms of computer time, has the advantage of being computationally robust and ensuring maximization of the QSM. The above-described algorithm and implementation using the Lamarckian GA with local simplex optimization will further be called the Quantum Similarity Superposition Algorithm (*QSSA*).

Implementation. Prior to the discussion of the results obtained with QSSA, some more details on the QSSA implementation should be discussed. QSSA starts by reading the Cartesian coordinates for all atoms in the different molecules. After this input the necessary ASA parameters are read, and all terms that are alignment independent in the 4-fold summation in eq 12 are calculated and stored in memory.

After this step, a collection of random sets of alignment parameters is chosen, and the QSM is calculated for each set. For each set, the QSM is locally optimized using the simplex method. Usually 100 initial random sets are constructed. After all these sets have been optimized with the simplex method, a genetic algorithm population is constructed from these 100 sets. The genetic algorithm is then started, where intermediately, between all generations; the simplex method is used as a local, intermediate optimizer. The genetic algorithm is a fairly standard implementation using a roulette wheel selection scheme but with a somewhat higher rate of new specimen entering the population. This is done to make sure that the genetic algorithm can explore a large range of possible sets of alignment parameters. Experience shows that usually only a limited number of generations should be produced, 20 generations usually being sufficient.

As will be described below, the TGSA produced alignment can sometimes be used as a member of the first set of normally random chosen sets of parameters. This is simply done to benefit from those instances where TGSA and QSSA produce rather similar alignments. In that case, using TGSA can help introduce a good alignment already in the first generation of the genetic algorithm. It should be stressed that, although sometimes beneficiary, there is no need to include the TGSA alignment in the first set of parameters. QSSA works perfectly without a preliminary TGSA step. In fact, in the applications discussed below, nearly all optimal

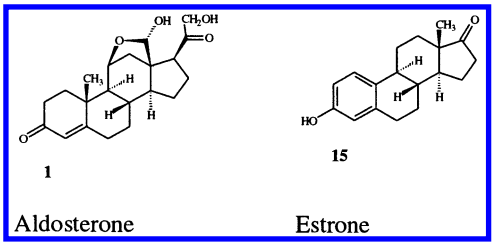

Figure 1. Structures of aldosterone and estrone used in the comparison of TGSA and QSSA alignments and molecular quantum similarity.

QSSA alignments were produced from other alignments than the TGSA alignment.

QSSA can be used for any set of quantum objects. In the present case we will solely consider molecules as quantum objects. It does not matter for QSSA whether these molecules show some degree of congener nature or not. Congener refers in this context to molecules which in a logical visual context show structural similarity or a similar origin. The steroids considered below form a congener set, whereas the binding isomers of the same stoichiometry form a globally noncongener set, although there may be some congener character between two molecules in the set. From the QSSA derived matrix **Z**, similarity dendrograms may be obtained using the techniques described previously by Bultinck et al.[25] More specifically, Carbó similarity index transformed matrices **Z** were used to obtain the dendrograms shown below.

APPLICATION AND DISCUSSION

To test the alignment approach presented above, two examples are addressed in detail. In the first, a set of 31 globulin binding steroids, as used previously by Cramer et al.,[26] Wagener et al.,[27] and Carbó et al.,[28,29] is considered. Bultinck et al.[25] used this same set of steroids in the development of a dendrogram based method for molecular clustering, where also the molecular similarity as defined in eq 2 is used. The clustering based on TGSA aligned structures was previously examined.[29]

The second example consists of a set of organic molecules, all binding isomers with the common stoichiometry $C_4H_6O_2$.

In the two cases, TGSA maximized molecular similarities were calculated using ASA densities, and the TGSA alignments were used as one member of the first generation in the QSSA method. From the results it appeared that only in very few cases, TGSA alignments were not improved by the QSSA method. In most cases, the optimal QSSA alignment was even not produced from the TGSA alignment but rather originated from a different set of alignment parameters.

Steroids. Ninety-nine random sets of alignment parameters were created to form a first generation in the Lamarckian QSSA algorithm. These 99 sets were augmented with the optimal alignment from TGSA. It was found that taking into account explicitly the electron density of both molecules can yield an alignment other than TGSA and even a relevant difference in the molecular similarity measure. An outspoken case lies in the similarity between molecules 1 and 15 of the set of steroids. The molecular structures are given in Figure 1.

Both molecules originate from the basic steroid framework and can be considered as congener molecular species. TGSA alignment was performed, and quantum similarity was calculated. This revealed a TGSA based value of 224.54 using the previously published ASA parameters. When using the present scheme, a slightly different alignment is produced, and the resulting maximum similarity found amounts of 312.52. This is naturally a relevant difference, amounting to almost 40%. After performing the QSSA calculation of the quantum similarity matrix for the entire steroid set, no uniform increase in the maximum value of Z_{AB} for all molecules was found. As an example, on average over all values of the first column of the quantum similarity matrix, the relative difference between the TGSA based similarity and the QSSA based one, amounts to some 14%, the present algorithm giving higher similarities. Considering the entire collection of molecules, one finds that the smallest differences between TGSA and the present algorithm are usually found for those molecules which differ little in molecular structure, e.g., the addition of a single methyl group. For molecules that differ more, TGSA performs clearly less good, and more outspoken differences start to occur. For the molecules which differ only slightly in molecular structure, in the present application it is found that QSSA yields almost no improvement in the alignment, whereas in the other cases, QSSA produces different alignments with appreciably higher similarities. This not only shows the sensitivity of the similarity measure versus the alignment procedure but stresses the necessity to go beyond a topographical approach to maximize the quantum similarity. Since there is no systematic increase in the QSM when using QSSA versus TGSA, for some elements of the quantum similarity matrix **Z** one may obtain a similarity with TGSA that differs little from the QSSA one, but for others the values can differ by 40%. This may naturally have an important influence on the ranking of molecules according to similarity and derived relations such as in quantum QSAR.

Binding Isomers of $C_4H_6O_2$. Again, it should be stressed that not only QSSA succeeds at finding alignments that yield higher values for the molecular quantum similarity measures but also that the new maximal similarity alignment is not necessarily produced by starting from the optimal TGSA alignment. An experiment was carried out to test whether omitting the TGSA alignment from the initial population in QSSA has any consequences on the final QSSA **Z** matrix. It was found that the final **Z** matrix was no different, indicating that the TGSA alignment prior to QSSA is not a necessary condition for the good functioning of QSSA. For the diagonal elements Z_{ii}, constituted by the quantum self-similarity measures, no alignment step is necessary, and so these elements remain unchanged when using TGSA or QSSA for the construction of **Z**. The other elements can, however, change drastically as described above. This means that one does not obtain some kind of simple shift in every element of the matrix but that a fundamentally different **Z** matrix is obtained. This behavior is even more apparent, when constructing an alternative **Z′** matrix, using Carbó indices[20,29] or stochastic transformations.[29,30]

The better performance of QSSA compared to TGSA may be illustrated by the example of $C_4H_6O_2$ binding isomers. A total of seven molecules was chosen. All of these are commercially available compounds,[31] and their Lewis structures are shown in Figure 2.

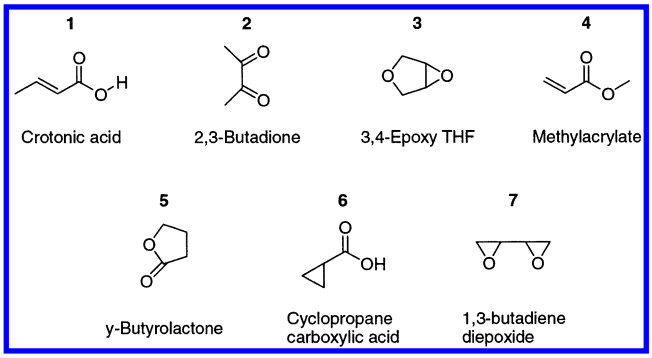

Figure 2. Seven binding isomers of $C_4H_6O_2$ used in the molecular alignment study.

These molecules show different extents of congener character among each other. TGSA alignment is expected to perform very good in the case of e.g., crotonic acid versus methylacrylate, since there are apparent several common structural features. Other molecules do not show such obvious common substructure elements. TGSA and QSSA alignments were performed, with the resulting 7×7 symmetrical similarity matrices:

$$\mathbf{Z}_{\mathrm{TGSA}} = \begin{bmatrix} 293 & 136 & 78 & 116 & 180 & 225 & 76 \\ & 293 & 65 & 175 & 132 & 87 & 84 \\ & & 293 & 98 & 165 & 122 & 104 \\ & & & 293 & 197 & 214 & 82 \\ & & & & 293 & 143 & 67 \\ & & & & & 293 & 107 \\ & & & & & & 293 \end{bmatrix}$$

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

It is obvious that the similarity matrices may exhibit large differences. Care should be taken in considering a higher QSM Z_{AB} compared to a value Z_{CD} as indicating a higher similarity between A and B than between C and D. The degree of similarity is better expressed using Carbó indices, where perfect similarity corresponds to a value of 1, and all other cases are expressed in the interval]0,1]. In the present example, due to the fact that all diagonal elements are nearly the same, there is a relation between higher extents of similarity and elements of **Z**, but care should be taken not to consider this as a general feature. Differences between the two **Z** matrices may be quantitatively measured in global terms, for instance, by computing the Euclidean distance between both matrices, which produces the value: 319.6. Also, the cosine of subtended angle computed in Euclidean metric produces a value of 0.977. However, as commented in a recent paper, mathematical constructs such as similarity matrices may be better studied within an appropriate vector semispace metric[32] based on Minkowski's metric. In this case the corresponding values are 100.5 for the distance and 0.993 for the cosine of the subtended angle. Such results indicate that distances in either formulation have better discriminating

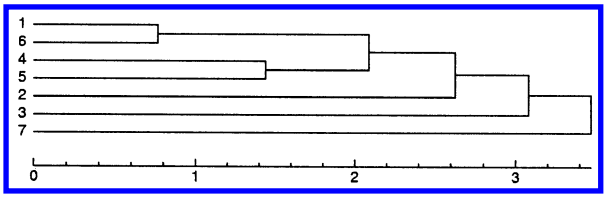

Figure 3. Molecular quantum similarity based dendrogram using TGSA alignment.

features than cosines, supporting the claim of the present work that both matrices exhibit different numerical contents. For some combinations of molecules, TGSA succeeds at finding an alignment that cannot be improved to a great extent by QSSA. In most cases, however, QSSA does give appreciably higher similarities. The most remarkable example is that of molecules 1 (crotonic acid) and 4 (methylacrylate). Using QSSA, the quantum similarity measure is nearly double the TGSA based one, although inspection of the structures would have indicated good performance of TGSA. This clearly illustrates that QSSA is preferable over TGSA, even for molecules that are expected to be structurally similar. The observed differences are related to the fact that in the QSSA procedure one aims at finding the alignment that maximizes the molecular similarity. In TGSA one tries to align molecules in such a way that they maximally align similar structural elements. In the case of congener molecules, QSSA will also produce an alignment where a maximum number of atoms coincide, and due to the congener nature of the molecules involved, many of the coinciding atom pairs will involve the same chemical elements. When the congener nature is lost, however, TGSA will still try to align a maximum number of atoms of the same element in both molecules. The present example also indicates that, even for relatively congener molecules, one needs to be careful in using TGSA, and TGSA results should always be confronted against a QSSA test. It is clear that from a quantum similarity point of view, QSSA is preferable, since one finds a unique upper limit of the quantum similarity measure. The fact that the self-similarities differ so little for the different molecules is due to the fact that all molecules have the same stoichiometry, and the use of ASA makes the electron density a sum of atomic terms.

It is immediately clear that in the field of Quantum QSAR, a different matrix **Z** will have an appreciable effect on the SAR. This is because in Quantum QSAR, one uses rows of **Z** as descriptors for the molecule within the space spanned by the *N* molecules composing the matrix. If these descriptors change, sometimes appreciably, due to the use of a different alignment scheme, one can immediately expect a different resulting QSAR model. The QSSA procedure is in this context preferable, since it gives a unique upper limit to the maximal quantum similarity, whereas similarities based on topographical considerations may very well differ depending on which structural criteria are used.

A related field where the different alignment procedure may have important consequences is that of molecular clustering. A quantum similarity based scheme for the construction of quantum object dendrograms was described previously by Bultinck et al.[29] Based on the QSM matrix obtained using TGSA or QSSA, one will obtain different dendrograms, as is shown in Figures 3 and 4 for the set of

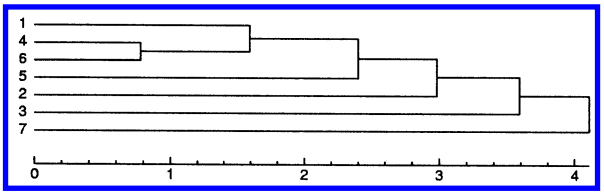

Figure 4. Molecular quantum similarity based dendrogram using QSSA alignment.

binding isomers where the Carbó index transformation technique for clustering has been used.[25]

At first glance the dendrograms are similar, but there is an important difference. Figures 3 and 4 clearly indicate that it is important to use QSSA as an alignment procedure. Looking at the matrices **Z** derived using TGSA and QSSA, one of the most obvious differences lies in the element Z_{14} as discussed above. The element Z_{16} is nearly the same in both matrices. When using QSSA both elements Z_{14} and Z_{16} are also nearly identical. Interestingly, the element Z_{46} in QSSA has grown quite large and is now even the largest nondiagonal element in the entire matrix. Since all diagonal elements are almost exactly equal, the values of the off-diagonal elements of **Z** already reflect the similarity without Carbó index transformation. It is thus found that the most similar pair of molecules when using QSSA consists of molecules 4 and 6. Molecule 1 is then picked up in the cluster of molecules 4 and 6. It is clear not only that the similarity between molecules 1 and 4 was strongly underestimated in TGSA but also that it failed to recognize the most similar pair.

Even in this small example, one finds a relevant impact of using QSSA instead of TGSA. When considering bigger sets of molecules, differing in stoichiometry and showing outspoken noncongener character, experience shows that the differences between $\mathbf{Z}_{TGSA}$ and $\mathbf{Z}_{QSSA}$ may grow quite large, and the resulting dendrograms can differ to large extent.

Regarding the information content of the QSSA and TGSA alignments and the associated values for the molecular quantum similarity measures, the following should be noted. Different topogeometrical methods can differ in many aspects, such as their expressions and algorithms used to express chemical structure, their alignment algorithm, and possible thresholds used in the algorithms. Different topogeometrical methods can as a result easily yield different optimal structural alignments. When the QSM is then calculated, different QSM values are likely to occur. This is not the case with QSSA. Since QSSA aims at finding the global maximum of the QSM, and assuming the nondegenerate character of the global maximum, only one unique upper limit QSM can be found. From a mathematical point of view such an approach is much more consistent with the idea of quantum similarity, since it takes away several ambiguous aspects that are present when one uses structural alignments. The aspect of consistency in the present study means that both the alignment and the QSM originate from the same source, namely the electron density. A larger value for some element Z_{AB} in **Z** when using QSSA compared to TGSA is always better, since it means that one approaches better the global maximum of the QSM, which is a unique upper limit. It should be noted that it is not unimaginable that the TGSA based values in the **Z** matrix could give a better link to an observed behavior toward some receptor or to some observed bioactivity than the QSSA based matrix. This would not allow one to conclude that TGSA is a better technique. Consistency is a more important consideration than the fact that a model gives some interesting result. In fact, it can be shown that structural alignments in QSM studies yield quite strong violations of internal distance geometry considerations.[33]

The algorithm presented above can be used throughout for QSM expressions like eq 2, whatever the origin of the electron densities. Naturally, all elements should be obtained using the same level of method to obtain these electron densities. The QSSA obtained matrix **Z** should always be interpreted keeping in mind the actual method used to obtain the electron densities. In the present study, approximate electron densities are used. Naturally, more precise electron densities as obtained using some higher level quantum chemical method would be more appropriate. The QSSA algorithm does, however, require quite a number of QSM evaluations with different orientations of the molecules in space. Taking for instance Hartree−Fock expressions for the electron density, the electron density (2) becomes[15]

$$\rho(\mathbf{r}) = \sum_{\gamma}\sum_{\nu} P_{\gamma\nu}\phi_{\gamma}(\mathbf{r})\phi_{\nu}^{*}(\mathbf{r}) \qquad (14)$$

where $P_{\gamma\nu}$ is the appropriate element of the density matrix and ϕ are the basis functions. It is then immediately clear that for a QSM definition as in eq 2, a very large amount of integrals have to be evaluated. The basis functions themselves are usually contractions of Gaussian type orbitals of different types, thereby making the integrals rotationally variant and also harder to solve than in case of the ASA use of s-type orbitals only. The fact that already in one of the smallest ab initio models, namely Hartree−Fock, the integrals are more difficult, larger in number, and no longer rotationally invariant, together with the fact that QSSA requires repeated evaluation of the QSM, makes QSSA with ab initio densities not yet feasible. The algorithm itself will, however, remain the same if more accurate electron densities become feasible in molecular QSSA alignment.

CONCLUSIONS

A new approach is described to align molecules on the basis of molecular electron density functions. This approach contains the intuitively felt link between, on one hand, molecular similarity and molecular alignment on the other. The molecular similarity is hereby calculated through the overlap integral of molecular densities between two molecules, and the alignment is performed in such a way as to maximize the molecular similarity. This maximization is performed in terms of the relative orientation of the second molecule's coordinate system with respect to that of the first molecule, using three translation parameters and three Euler angles. Given the large number of local maxima, this optimization involves a Lamarckian genetic algorithm with the simplex method as a local optimizer.

Next to having a sounder quantum chemical basis, the present technique also performs better in cases where the topography of the molecules differs appreciably. In such cases QSSA still can be used without problems, and it has been shown that even for relatively congener molecules, the present technique may yield optimal values for the overlap

similarity measure. Experience shows that even for relatively congener molecules, QSSA should be applied to check whether TGSA has performed adequately.

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 Polytechnic 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. X. Gironés wishes to acknowledge the *University of Girona* for a predoctoral fellowship. The authors are very grateful to Dr. L. Chen for his expert advice.

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) Girones, X.; Robert, D.; Carbo-Dorca, R. TGSA: A molecular superposition program based on topo-geometrical considerations. *J. Comput. Chem.* **2001**, *22*, 255−263.
(5) 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.
(6) Carbó-Dorca, R.; Besalú, E. A general survey of Molecular Quantum Similarity. *J. Mol. Struct. (THEOCHEM)* **1998**, *451*, 11−23.
(7) Carbó-Dorca, R.; Robert, D.; Amat, L.; Gironés, X.; Besalú, E. Molecular Quantum Similarity in QSAR and Drug Design. *Lecture Notes Chem.* **2000**, *73*.
(8) 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.
(9) 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.
(10) Besalu, E.; Girones, X.; Amat, L.; Carbo-Dorca, R. Molecular quantum similarity and the fundamentals of QSAR. *Acc. Chem. Res.* **2002**, *35*, 289−295.
(11) 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.
(12) 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.
(13) Amat, L.; Carbo-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.
(14) 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.
(15) *Modern Quantum Chemistry*; Szabo, A., Ostlund, N. S., Eds.; Dover Publications: New York, 1996.
(16) ASA coefficients used in the present work may be found on the Internet at http://iqc.udg.es/cat/similarity/ASA/funcset.html.
(17) Girones, X.; Amat, L.; Carbo-Dorca, R. Modeling large macromolecular structures using promolecular densities. **2002**, *42*, 847−852.
(18) Constans, P.; Amat, L.; Carbó-Dorca, R. Toward a Global Maximization of the Molecular Similarity Function: Superposition of Two Molecules. *J. Comput. Chem.* **1997**, *18*, 826−846.
(19) *The CRC Concise Encyclopedia of Mathematics*; Weisstein, E. W., Eds.; CRC Press: 1998.
(20) Carbó, R.; Besalú, E.; Amat, L.; Fradera, X., On Quantum Molecular Similarity Measures (QMSM) and Indices (QMSI). *J. Math. Chem.* **1996**, *19*, 47−56.
(21) Bultinck, P.; Langenaeker, W.; Lahorte, P.; De Proft, F.; Geerlings, P.; Van Alsenoy, C.; Tollenaere, The electronegativity equalization method I: parametrization and validation for atomic charge calculations *J. Phys. Chem. A* **2002**, *106*, 7887−7894.
(22) Bultinck, P.; Langenaeker, W.; Lahorte, P.; De Proft, F.; Geerlings, P.; Waroquier, M.; Tollenaere, J. P. The electronegativity equalization method II: applicability of different atomic charge schemes. *J. Phys. Chem. A* **2002**, *106*, 7895−7901.
(23) *Molecular Modelling, Principles and Applications*; Leach, A. R., Ed.; Prentice Hall: Harlow, UK, 2001.
(24) Bultinck, P. Analytical expressions for gradients and Hessians in molecular superposition. Unpublished material.
(25) Bultinck, P.; Carbó-Dorca, R. Molecular quantum similarity matrix based clustering of molecules using dendrograms. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 170−177.
(26) 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.
(27) 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.
(28) 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.
(29) 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.
(30) 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.
(31) The $C_4H_6O_2$ bonding isomers were chosen from the Acros Chemical Catalogue, which can be consulted via http://www.acros.be.
(32) Carbó-Dorca, R. Shell Partition and Metric Semispaces: Minkowski Norms, Root Scalar Products, Distances and Cosines of Arbitrary Order. *J. Math. Chem.* **2003**, *32*, 201−223.
(33) Bultinck, P.; Van Alsenoy, C.; Carbó-Dorca, R. Quality of approximate electron densities and internal consistency of molecular alignment algorithms in molecular quantum similarity. *J. Chem. Inf. Comput. Sci.*, Accepted for publication, 2003.

CI0340153