Communications on Quantum Similarity, Part 3: A Geometric-Quantum Similarity Molecular Superposition Algorithm

RAMON CARBÓ-DORCA, EMILI BESALÚ, LUZ DARY MERCADO
Institut de Química Computacional, Universitat de Girona, Girona 17071, Catalonia, Spain

Received 20 May 2010; Revised 2 July 2010; Accepted 3 July 2010
DOI 10.1002/jcc.21644
Published online in Wiley Online Library (wileyonlinelibrary.com).

Abstract: This work describes a new procedure to obtain optimal molecular superposition based on quantum similarity (QS): the geometric-quantum similarity molecular superposition (GQSMS) algorithm. It has been inspired by the QS Aufbau principle, already described in a previous work, to build up coherently quantum similarity matrices (QSMs). The cornerstone of the present superposition technique relies upon the fact that quantum similarity integrals (QSIs), defined using a GTO basis set, depend on the squared intermolecular atomic distances. The resulting QSM structure, constructed under the GQSMS algorithm, becomes not only optimal in terms of its QSI elements but can also be arranged to produce a positive definite matrix global structure. Kruskal minimum spanning trees are also discussed as a device to order molecular sets described in turn by means of QSM. Besides the main subject of this work, focused on MS and QS, other practical considerations are also included in this study: essentially the use of elementary Jacobi rotations as QSM refinement tools and inward functions as QSM scaling methods.

© 2010 Wiley Periodicals, Inc. J Comput Chem 00: 000–000, 2010

Key words: molecular superposition (MS); quantum similarity (QS); QS matrices; QS integrals; Carbó QS index; QS Aufbau principle; geometric QS MS algorithm; Kruskal trees; restricted elementary Jacobi rotations; inward functions of QS matrices

Introduction

While constructing a quantum similarity (QS) matrix (QSM) using any literature described process,1–7 the procedure internal coherence and the resultant matrix structure has to be more or less loosely related to the technique, formerly described as the QS Aufbau principle (QSAP). Such a principle has been recently proposed8 as an inherent property, which must be attached to QSM mathematical properties. The QSM coherence problem has been previously studied from another point of view, based on Euclidian QS distances.9 Neither the coherent construction of QSM nor its relationship with molecular superposition (MS) problems has been frequently studied, unlike other related theoretical problems associated with QS (see Part I of this series10 for an exhaustive bibliographic account of QS). In addition, the problem is far from completely solved: this is due to the optimal search complexity, inherent to the QS integrals (QSI) computational structure in position space.

On the other hand, since the first appearance of QSI, MS has obviously been a ubiquitous problem in the calculation of optimized QSI, see for example, refs. 11–16. However, this work will discuss by means of a relatively simple procedure that there is a possible way to obtain coherently constructed QSM, while taking into account at the same time optimal MS.

Hence, this article aims to describe in a detailed manner several facets of this possibility. The foundation of such a molecular MS procedure relates to the geometric properties of the intermolecular distance matrix and the computationally connected QSI. Because of the linking of the proposed algorithm, both with the geometric properties of the molecular atomic positions in space and the structure of QSI, the whole procedure could be named in a self-descriptive way as the geometric-quantum similarity molecular superposition (GQSMS) algorithm.

The information contained in QSM can be used in several ways. One of these possibilities can be associated to molecular set ordering, see, for example, refs. 1–7. Kruskal minimum spanning trees (KMSTs) have been used since the origin of QS. KMSTs make available a schematic way to visualize the relationships between molecular set elements, described with QSM.

Correspondence to: R. Carbó-Dorca; e-mail: quantumqsar@hotmail.com
Contract/grant sponsor: Spanish Ministerio de Educación y Ciencia; contract/grant number: CTQ2006-04410/BQU
Contract/grant sponsor: Spanish Ministerio de Ciencia e Innovación; contract/grant number: CTQ2009-09370

© 2010 Wiley Periodicals, Inc.
Therefore, they will also be discussed here to complete the application framework, provided by the GQSMS algorithm.

Any method connected to QSI optimization has an interesting common characteristic, though it leaves the diagonal elements of any QSM invariant. Such a computational peculiarity can be used to refine QSM in a way related to elementary Jacobi rotations. Here, we will also study how off-diagonal elements of the QSM can be augmented via a similarity matrix transformation using a restricted EJR procedure, leaving the involved diagonal elements invariant. Finally, the problem of scaling QSM will be briefly considered; there are many possibilities that can provide the invariants of matrices concept.

The whole set of algorithms and computational procedures described in this work is illustrated with a sequence of seven FORTRAN 95 (F95) standalone programs: the Molecular QS Program Suite (MQSPS). They can be downloaded from a specific web site created for this purpose, as has been customary in the previous studies Parts I and II of this Communications on QS series. The necessarily selective results included in this article have been directly obtained from the use of these aforementioned MQSPS codes. Every program folder containing the source codes will merit here a brief description. Additionally, several input–output information examples will be contained in every web MQSPS folder. The amount of information one can obtain from the provided programs is large enough to make it impractical to provide it in full within this study. Therefore, only a minimal part of such information will be specified in printed form for some chosen sample molecular sets; the rest of the provided input–output information can be easily retrieved from the program folders or created at user will from the MQSPS codes.

This work, according to these previous considerations, will be essentially organized in several parts as follows:

1. An introduction will first be given on the basic underlying theoretical and mathematical principles. The role of the intermolecular squared Euclidian distances matrix will be studied next: there, it will be shown how the complete sum of the elements of such a matrix can be expressed in a compact fashion and how such a simple form will become minimal, representing at the same time a plausible MS when reaching a minimum. This last finding will provide the pathways to the third and fourth sections, where an appropriate molecular alignment and the GQSMS algorithms will be described.

2. Next, the role of the KMST algorithm will be discussed, as a tool to order a given molecular set from the knowledge of the QSM, via the Carbo QS index matrix (CSIM). Presented as a QS ordering simple device, at the dawn of QSI theoretical development, the KMST algorithm was described and used for the first time as a MO ordering procedure in previous work. For the sake of completeness and better understanding of this work, the KMST procedure is given later on, in the form of F95 code shown in Appendix A.

3. A new concept, derived from KMST: the Kruskal tree ordering (KTO), will also be described in the next step as a tool to determine the order of the items contained within a known molecular set. This ordering is used to obtain the so-called Frozen Aufbau QSM (FAQSM), a collection of techniques to build QSM, which provide a unique set of coordinates for every element of the underlying molecular set. Such a unique set of coordinates ensures in a general manner the positive definite structure of QSM. From the KTO concept, other possibilities, mainly associated with circular and random permutations of the molecular indices, are described. They can be considered as an alternative set of algorithms to obtain unique collections of molecular coordinates and thus positive definite QSM.

4. Continuing with the development of this work, we discuss the possible QS refinement using EJR and the scaling possibilities furnished by inward functions of matrices, because two independent programs dealing with these two subjects are included in the MQSPS.

5. A next step in the computational framework of this article will deal in a paragraph containing a brief description of the source programs developed in this work, which are available in the web, see the ref. 28. The MQSPS codes are written in F95 and constructed to test the procedures described in this work. This section closes the theoretical–computational part of the article.

6. Finally, a small set of applications will be described. The provided examples involve several molecular sets, in addition to the so-called Cramer set of steroids, which has earlier been subject of several calculations related with QS in our laboratory. This application section will be kept to a minimal size, as it has been previously commented that the MQSPS can provide a large amount of information, which can be easily extracted directly from the web site and by using the programs it contains.

To provide the readers with as much information as possible on the subject of MS and the MQSPS structure, several appendices, besides that providing a KMST algorithm, are also included to offer other interesting source listings, contained within the web programs.

### Basic Principles

The MS problem related to QS computational questions has been previously studied from several points of view. Among others, associated with this laboratory work, can be quoted first a procedure purely based on QS and in addition the so-called topo-geometrical algorithm (TGSA). Furthermore, it is related to the more recent MS work of Bultinck et al. in refs. 15 and 16.

In any MS procedure involving a set of at least two molecules, which can be called \( \{A,B\} \), say, one can supposedly consider as unmoving, the molecule \( A \), whereas the other, the molecule \( B \), is submitted to translations and rotations (TR) of the coordinates associated with the fixed atomic positions. This TR manipulation can go on until some optimal MS is reached, according to a specific prefixed criterion, which can be a maximal QSI value, for instance. The previous references used such a schematic MS background procedure.

However, in principle, to obtain some optimal MS, an algorithm may be searched for, where both molecules are independently translated to a common origin and rotated afterward. These geometrical operations will be mentioned, as before, by TR of the molecular Cartesian coordinate frames. The advantage of such a TR process involving both molecules is that it can be applied, not only to a pair of molecular Cartesian frames but also simultaneously on an indefinite number of molecular structures. In this...
article, though, the procedure description and programing will remain bimolecular, whereas forthcoming work will deal with the problem of superposing simultaneously several molecular frames.

In fact, a bimolecular QSI, weighted by a positive definite operator, $\Omega(r,s)$ can be formally written (see, for example, refs. 36–43) as follows:

$$Z_{AB}[\Omega] = \langle \rho_A \Omega \rho_B \rangle = \int_D \rho_A(r) \Omega(r,s) \rho_B(s) dr ds.$$  \hspace{1cm} (1)

Alternatively, seeking for lesser time-consuming computer performance, one can use the simpler overlap QS measure:

$$\Omega(r,s) = \delta(r - s) \rightarrow Z_{AB} = \langle \rho_A \rho_B \rangle = \int_D \rho_A(r) \rho_B(r) dr.$$  \hspace{1cm} (2)

It is a well-known fact that the ultimate QSI value between GTO basis functions depends on the intermolecular atomic distances squared; see for more information, the first\(^{10}\) and second article of this series\(^{18}\) or any study of GTO integrals, like those of refs. 44 and 45, for example. This is the consequence of a general property, which can be associated to the analytical formulae attached to quantum mechanical integrals, based on GTO basis sets involving Hermitian operators.\(^{44–47}\)

As a simple example, which can serve to illustrate the idea, one can formally write

$$\Omega(r,s) = \delta(r - s) \rightarrow Z_{AB} = \langle \rho_A \rho_B \rangle = \int_D \rho_A(r) \rho_B(r) dr.$$  \hspace{1cm} (2)

To continue developing a MS algorithm essentially associated to the SqDM, it is interesting to describe the form of the SqDM complete sum as shown above in eq. (4), but now using explicitly the pair of molecular coordinates: $\{R^A = 1, N_A\}$ and $\{R^B = 1, N_B\}$ associated with both involved molecular frames. One can write the complete sum for SqDM:

$$\langle D^{(2)}_{AB} \rangle = \sum_{I=1}^{N_A} \sum_{J=1}^{N_B} (D^{AB}_{IJ})^2 > 0 \hspace{1cm} (4)$$

and the properties of this expression are relevant for MS application purposes.

As a simple example, which can serve to illustrate the crucial role assumed by the SqDM in MS, imagine, for instance, that both considered molecular frames $A$ and $B$ correspond to the same molecule in exactly the same geometrical configuration.

Therefore, in these quite particular circumstances, when both molecular structures just superpose, as a result, the complete sum of the above defined SqDM: $D^{(2)}_{AB}$ elements acquire its minimal positive definite value.

To continue developing a MS algorithm essentially associated to the SqDM, it is interesting to describe the form of the SqDM complete sum as shown above in eq. (4), but now using explicitly the pair of molecular coordinates: $\{R^A = 1, N_A\}$ and $\{R^B = 1, N_B\}$ associated with both involved molecular frames. One can write the complete sum for SqDM:

$$\langle D^{(2)}_{AB} \rangle = \sum_{I=1}^{N_A} \sum_{J=1}^{N_B} (D^{AB}_{IJ})^2 > 0 \hspace{1cm} (4)$$

and the properties of this expression are relevant for MS application purposes.

As a simple example, which can serve to illustrate the crucial role assumed by the SqDM in MS, imagine, for instance, that both considered molecular frames $A$ and $B$ correspond to the same molecule in exactly the same geometrical configuration.

Therefore, in these quite particular circumstances, when both molecular structures just superpose, as a result, the complete sum of the above defined SqDM: $D^{(2)}_{AB}$ elements acquire its minimal positive definite value.

To continue developing a MS algorithm essentially associated to the SqDM, it is interesting to describe the form of the SqDM complete sum as shown above in eq. (4), but now using explicitly the pair of molecular coordinates: $\{R^A = 1, N_A\}$ and $\{R^B = 1, N_B\}$ associated with both involved molecular frames. One can write the complete sum for SqDM:
\[
\left\langle D_{AB}^{(2)} \right\rangle = \sum_{j=1}^{N_A} \sum_{j=1}^{N_B} \left( \langle R^a_j | R^b_j \rangle + \langle R^b_j | R^a_j \rangle - 2 \langle R^a_j | R^b_j \rangle \right)
\]
\[
= N_B \sum_{j=1}^{N_A} \langle R^a_j | R^a_j \rangle + N_A \sum_{j=1}^{N_B} \langle R^b_j | R^b_j \rangle - 2 \sum_{j=1}^{N_A} \sum_{j=1}^{N_B} \langle R^a_j | R^b_j \rangle
\]
\[
= N_B \sum_{j=1}^{N_A} |R^a_j|^2 + N_A \sum_{j=1}^{N_B} |R^b_j|^2 - 2 \left( \sum_{j=1}^{N_A} |R^a_j| \sum_{j=1}^{N_B} |R^b_j| \right) \quad (5)
\]

At this stage, when defining a new pair of vectors as sums of the total atomic coordinates at each molecular frame, for example, using the following notation:
\[
R^A = \sum_{j=1}^{N_A} R^A_j, \quad R^B = \sum_{j=1}^{N_B} R^B_j,
\]
then, the complete sum of the SqDM (5) can be finally written as the simple expression:
\[
\left\langle D_{AB}^{(2)} \right\rangle = N_B \sum_{j=1}^{N_A} |R^A_j|^2 + N_A \sum_{j=1}^{N_B} |R^B_j|^2 - 2 \left( \sum_{j=1}^{N_A} R^A_j \sum_{j=1}^{N_B} R^B_j \right) \quad (6)
\]

The first two terms of the SqDM complete sum, as expressed in eq. (6), correspond, respectively, to summations of rotationally invariant Euclidean norms of the atomic coordinate vectors, involving each molecule separately. However, one must be aware that norms can suffer variations when performing translations.

The third element of eq. (6) (5) can also be considered as a scalar product of the vector sums of the atomic coordinates of each molecule. For a known fixed pair of atomic coordinate sets in molecules A and B, the scalar product \( R^A \cdot R^B \), when becoming maximal, will obviously minimize the complete sum of the intermolecular squared distances. Thus, sets of TR operations performed on the involved vectors, tending to produce a maximum \( \langle R^A \cdot R^B \rangle \) value, will tend to produce a minimum value of \( \left\langle D_{AB}^{(2)} \right\rangle \).

Proceeding in this manner, a better MS, in the sense of optimized according to this criterion, will certainly appear because of the application of this algorithm.

A Preparatory Algorithm

According to this trivial property, which can be attached to any bimolecular SqDM, first a set of translations assigns both molecules a common origin. Then, convenient rotations on both vector pairs \( \{ R^A, R^B \} \) performed separately\(^1\), but bringing both vector sums to any arbitrarily chosen coincident space direction (any three-dimensional direction or axis can be a convenient choice, but here the \( x \)-axis direction, for instance, has been chosen) will certainly maximize the scalar product \( R^A \cdot R^B \). This yields a minimum of the complete sum of the SqDM.

\(^1\)Practical reasons have compelled us to use translations and the vector coordinate sums over the nonhydrogen atoms present in the associated molecular structure.

Such an optimal orientation of the atomic coordinates can plausibly produce in turn a (or an almost near) maximal overlap QSI\(^2\).

Finally, when this collection of mathematical properties is assembled and calculated in full for a molecular pair, one perhaps can also take into consideration that this may indicate a better MS of such a pair of molecules. This can be visualized when all of a set of homogeneous molecules, treated under the GQSMS algorithm, is depicted.

Redirection of a Vector Toward the \( x \)-Axis

To set up the most simple, but certainly most important part of the procedure, we describe how to redirect any vector of the form, \( R = (r_1, r_2, r_3) \), in three-dimensional spaces in order that, at the end of the redirection manipulation, the vector points toward the \( x \)-axis direction, say.

This can be performed in two rotation steps. First, an initial rotation cosine and sine set is defined, using:
\[
c_a = r_1 (r_1^2 + r_2^2)^{-1/2} \quad c_s = r_2 (r_1^2 + r_2^2)^{-1/2}
\]
and second, another set cosine and sine of rotation is defined as follows:
\[
c_b = (r_1^2 + r_2^2)^{1/2} (r_1^2 + r_2^2 + r_3^2)^{-1/2} \quad s_b = r_3 (r_1^2 + r_2^2 + r_3^2)^{-1/2}
\]

The rotated final vector \( \mathbf{L} = (\lambda_1, \lambda_2, \lambda_3) \) components can be easily written when the auxiliary scalar is also previously defined:
\[
a = c_a r_1 + s_a r_2
\]
Then, the rotated vector components can be written as follows:
\[
\lambda_1 = c_b a + s_b r_3
\]
\[
\lambda_2 = -s_a r_1 + c_a r_2 = 0
\]
\[
\lambda_3 = -s_b a + c_b r_3 = 0
\]

Each two-step rotation cosines and sines, for the two summation vector pair \( \{ R^A, R^B \} \), can be easily obtained for each vector separately, using the previously described simple algorithm as given above. The same pair of two-step rotations can be applied over all the raw coordinates of each molecule.

Practical Considerations

Therefore, once the cosines and sines of the rotation angles are known, a well-oriented rotated vector can be obtained for every atomic coordinate vector in both molecules. The resultant rotated atomic coordinates\(^3\) will usually provide a minimal sum of the squared intermolecular distances.

The rotated atomic coordinates can be used afterward to obtain the corresponding QSI, which in most of the cases shall

\(^2\)One must we aware that although it may be possible that the attached QSI is maximized, it could not be proved that all possible cases will behave in this way. Therefore, any algorithm based on the squared distances matrix shall be constructed accordingly, taking into account that a test shall be present to confirm or not the associated QSI maximization.

\(^3\)Once obtained the rotations over the vector sums, which are associated only with heavy atoms, obviously the rotation shall be applied over the coordinates of all the atoms of the molecule, hydrogen atoms included.
possess a greater value than the one obtained by simple translation, superposing two atomic centers.\(^4\)

However, the warning must be repeated: before the described rotation scheme is performed, translations of both molecules to a common atomic origin must have already been performed. These a priori translations are needed to obtain a unique origin shared for both structures, before coherently performing the redirection rotations of both coordinate sum vectors. This has been done in practice by producing translations of the complete molecular coordinates set one at a time, excluding the hydrogen atoms.

A possible computational strategy is given by the following algorithm:

**Algorithm 1. Maximal value of a general QSI between two molecular structures**

Let: \(Z_{AB} = 0\)

**Loop** over atoms of molecule \(A\): \(I = 1, N_A\)

Translate the atomic coordinates of \(A\): \(\forall K \in A: (^{(1)}R^A_K \rightarrow R^A - R^A_J)\)

Compute the vector sum: \((^{(1)}R^A) = \sum_{K=1}^{N_A} (^{(1)}R^A_K)\)

Search for cosine and sine to rotate: \(L_A \leftarrow (^{(1)}R^A)\)

Rotate all coordinates of molecule \(A\): \(\forall K \in A: (^{(1)}R^A_K \leftrightarrow (^{(1)}R^A_K)\)

**Loop** over atoms of molecule \(B\): \(J = 1, N_B\)

Translate the atomic coordinates of \(B\): \(\forall L \in B: (^{(1)}R^B_L \rightarrow R^B - R^B_J)\)

Compute the vector sum: \((^{(1)}R^B) = \sum_{L=1}^{N_B} (^{(1)}R^B_L)\)

Search for cosine and sine to rotate: \(L_B \leftarrow (^{(1)}R^B)\)

Rotate all coordinates of molecule \(B\): \(\forall L \in B: (^{(1)}R^B_L \leftrightarrow (^{(1)}R^B_L)\)

Compute the QSI: \(\langle \rho \Omega \rho_B \rangle\), using the TR coordinates: \(\{L^A_{\ell}\} \setminus \{L^B_{\ell}\}\).

If \(\langle \rho \Omega \rho_B \rangle > Z_{AB}\) then: \(Z_{AB} \leftarrow \langle \rho \Omega \rho_B \rangle\)

**End Loop I**

**End Loop J**

**End of algorithm 1**

At the end of the previous algorithm, the optimal QSI has been computed for the particular molecular pair: \(Z_{AB}\), based on the previously described TR atomic coordinate scheme.

Such an algorithm has been implemented in one of the stand-alone programs associated with the present MQSPS and described below in the section Program Sequence Description.

\(^4\)It is also interesting to look at the possibility of refining the MS, as obtained in the way described until now, by rotating the atomic coordinates of either of the two molecules around the \(x\)-axis. This has been tested without any appreciable MS improvement, but at the cost of longer computer time. Thus, this \(x\)-axis extra rotation will not be further discussed in this article. Possible future improvements of the GQSMS algorithm will be left to future work.

---

**The Geometric Quantum Similarity Molecular Superposition Algorithm**

It must be noted that although the previous Algorithm 1 leads to a pairwise optimal molecular QSM: \(Z = \{Z_{PQ}\}\), and every QSM element \(Z_{PQ}\) is maximal in the sense discussed above, the QSM \(Z\) is not necessarily a positive definite matrix.

This is mainly because every molecular pair optimization may have manipulated each chosen molecule on a different TR frame, which may not necessarily be coincident with the frame of another molecular pair in the QSM \(Z\). A simple example can be given of this situation, which is difficult to envisage at first. Perhaps within the TR of the pairs \(AB\) and \(AC\), molecule \(A\) has been chosen to contribute to the optimal QSI \(Z_{AB}\) with a set of atomic coordinates, with a different origin and rotated in a different way from molecule \(C\), when the QSI \(Z_{AC}\) is optimized separately.

A coherent frozen Aufbau quantum similarity (FAQS) procedure will primarily consist in having for all the considered molecular pairs:

\[
\{(A; P)|A, P = 1, M|P \neq A\}
\]

a unique TR framework associated with every molecule \(A\) for all the rest of \(M - 1\) molecules in the studied set.

**Carbó Quantum Similarity Index Matrix of a Molecular Set**

To preserve the ideal goal as stated above, one can design a nonarbitrary fixed procedure applicable to any molecular set. Therefore, it is best to proceed first to apply Algorithm 1 to obtain a possibly nonsymmetrical QSM, \(Z\).

Although such a QSM could not be coherent from the QSAP point of view, it can be used safely to construct a CSIM\(^3\): \(C = \{C_{AB}\}\). Such a matrix is just the QSM with the elements divided by the square roots of the corresponding quantum self-similarity (QSS) integrals (QSSI), at the same time contained into the diagonal elements of the QSMIZ itself, that is:

\[
\forall A, B: C_{AB} = Z_{AB}(Z_{AA}Z_{BB})^{-\frac{1}{2}}.
\]

This is identical to defining the diagonal matrices:

\[
\Theta = \text{Diag}(Z) = \text{Diag}(Z_{II}|I = 1, M),
\]

and the inverse square root matrix:

\[
\Theta^{-\frac{1}{2}} = \text{Diag}\left((Z_{II})^{-\frac{1}{2}}|I = 1, M\right).
\]

Then, one can transform the original QSM according to the matrix algorithm:

\[
C = \Theta^{-\frac{1}{2}}Z\Theta^{-\frac{1}{2}} \rightarrow \text{Diag}(C) = I.
\]

Such a resulting matrix structure does not possess the different magnitudes of the QSI elements of the original QSM, but ensures that the elements of the CSIM \(\{C_{AB}\}\) become a set of
numbers lying in the interval \([0,1]\). That is:
\[
\forall A, B (A \neq B) : C_{AB} \in [0, 1].
\]

**Kruskal Minimum Spanning Tree and Kruskal Tree Ordering of a Molecular Set**

When known, the elements of the CSIM can be used to design an automatic procedure in order to choose a systematic reordering of the elements of any chosen molecular set. For this task of reordering the initial molecular set, there are a swarm of possible procedures, see, for example, ref. 22, including the so-called Kruskal’s Minimum Spanning Tree (KMST) Algorithm. The KMST algorithm has been earlier explicitly described and applied by one of us. In Appendix A, an improved F95 version of the KMST algorithm is given to facilitate understanding and implementation of the whole procedure.

The search for a KMST noncycling connection of the working molecular set elements first takes the most similar molecular pair. When known, the elements of the CSIM can be used to design an automatic procedure in order to choose a systematic reordering of the molecular set elements first takes the most similar molecular pair.

When the KMST is finally obtained, from the information contained within the tree branches, a new ordering of the molecular set is obtained: the Kruskal tree ordering (KTO). This constitutes a systematic reordering of the molecular set indices, valid in any circumstance and leaving all the QSM properties invariant.*

To obtain such a KTO from the KMST, an algorithm can be designed taking as the first molecule the one with the first index appearing in the KMST arrangement, the second as the one attached to the second nonrepeated index appearing in the KMST, and so on.

**An Illustrative Simplest but General Example of KTO Computation**

Suppose, for instance, three molecules initially and arbitrarily labeled as \(\{m_1, m_2, m_3\}\). Once a QSM and the associated CSIM are computed, suppose that the CSIM yields a KMST, which connects molecule 3 with molecule 1 and molecule 2 with molecule 1. This KMST can be represented just by a sequence of two molecular pairs as \(\{(m_3, m_1); (m_2, m_1)\}\).

It can also be graphically illustrated as the pictogram:

```
 3 ——— 1 ——— 2
```

Then, the KTO construction will yield in this case: \(m_3 \succ m_1 \succ m_2\). The symbol \(A \succ B\) just means: object \(A\) precedes object \(B\) or the equivalent relationship: object \(B\) follows object \(A\).

Appendix B includes a F95 code to obtain the KTO, given any KMST structure.

*Any permutation of the indices, arbitrarily attached to the background molecular set elements, leaves the QSM properties invariant.

A KTO obtained from any molecular set, previously ordered in an arbitrary fashion, ensures the subsequent reproducibility of the results, when QS procedures are applied to it.

This way of proceeding can be designed in the following way by using the next algorithm:

Algorithm 2. **Construction of a TR coherent FAQSM.**

0) Suppose the coordinates for a given molecular set are known:
\[
\{ R_k^n[p = 1:M]\}
\]
1) Use Algorithm 1 to construct an initial optimized QSM: \(Z\).
2) Compute the CSIM: \(C\) and use the KMST algorithm (Appendix A) to obtain a new KTO (Appendix B) molecular set index vector: \(m) = \{m_1 : l = 1:M\}\).
3) Define a logical vector with all elements false: \(\mu\).
   **Loop** \(P = 1, M - 1\)
   **Loop** \(Q = P + 1, M\)
   Choose the molecules according to the indices: \(m_P = A \land m_Q = B\):
   \[
   \Theta = \ominus(\delta_{A} \land \delta_{B} = F) \land \Theta)
   \]
   
   **if** \(\mu_i = F) \land \Theta) \land \Theta)
   **Compute** the atomic coordinates of \(A\):
   \[
   \forall K \in A; \quad R_A^K \rightarrow R_A^K = \sum_{K=1}^{N} R_A^K
   \]
   Search for cosine and sine to rotate all coordinates of molecule \(A\) yielding:
   \[
   \Theta_{A} = \Theta_{A}(R_A^K = \{R_A^K[K \in A}\)
   \]
   **end if**
   **else**: \(\Theta_{A} \omega_{L} = \omega_{L}\)

**end if**

**Compute** the QSI: \(\Theta_{AB} = \{\rho_{AB} \Theta_{B}\}\), using: \(\Theta_{A} \land \Theta_{B}\)
**end if**

\[\Theta > \Theta\) then: \(\Theta \leftarrow \Theta_{AB}\)

**if** \(\not= \Theta\) \land \Theta) \land \Theta)
**End Loop** \(J\); **End Loop** \(I\)

**if** \(\not= \Theta\) \land \Theta) \land \Theta)
**End Loop** \(J\); **End Loop** \(P\)

**End of Algorithm 2.**

Some Remarks on the Properties of Algorithm 2

The previous use of Algorithm 1 in Algorithm 2 is optional. It aims to obtain an initial QSM and the KTO for the involved
molecular set. Such an ordering will depend on the nature of the set itself, which will yield the appropriate QSM, and of the KMST algorithm. Thus, the resultant ordered molecular set will not depend on any arbitrary factor but only on the structure of the initial QSM.

However, as the properties of any QSM do not depend on the order of the attached molecular set, this starting manipulation could be omitted and any ordering of the molecular indices can be used. Nevertheless, it is interesting, from the methodological point of view, to have a molecular ordering leaving all the QS information invariant, and also somehow independent of the observer and reproducible, whatever initial order of the molecular set is chosen.

A standalone F95 program has been constructed to implement Algorithm 2 and included into the MQSPS. Some features of such a program are commented in the section Program Sequence Description below.

Extension of the Algorithm 2 to Triple Density QSM

The procedure associated to Algorithm 2 constitutes quite a general methodology, as it can be applied with almost the same structure to the coherent optimal computation of triple density QSM (TDQSM).\(^\text{48}\) Such QSIs are calculated, with three instead of two molecular density functions (DFs), as in the overlap QSI form shown in eq. (7), that is:

\[
Z_{\text{CAB}} = \int_D \rho_A(r)\rho_B(r)\rho_C(r)dr.
\]

The triple density quantum similarity integrals (TDQSI) defined above can be considered as QSI, where the positive definite operator has been chosen as a third DF. Algorithm 2 has to be slightly modified in this case, including two more loops within the four already present: one running over the molecules and the other over each atom in the additional molecule of the previous loop.

The same can be said for QSI involving multiple DFQSM construction. Every new DF, appearing in the corresponding multiple DFQSI elements, implies the addition of two more internal loops.

The structure of Algorithm 2 has the characteristic such that, used as described or in triple or higher QSM computations, it is completely parallelizable, at least in the section of QSI integrals within the four existing loops.

A forthcoming study will deal with theoretical, computational, and practical aspects of triple DF measures.

Circular Permutations of Molecular Indices Algorithm

Another possibility of the application of Algorithm 2 consists of using systematically the starting order of the indices, which correspond to the circular permutations of the canonical molecular indices.

Using the initial indices or the KTO obtained from the KMST information, one can compute \(M\) QSM, using the circular permutations of the given initial molecular indices. Accordingly, QSMs are coherently constructed, taking each index-reordered molecular structure in turn and optimizing the rest of the molecules, leaving their coordinates frozen for all the molecular pairs entering the QSI.

Every QSM of the set \(\{Z_{il} l = 1, M\}\) obtained in this way will be coherent, in the sense that every element for the \(l\)-th QSM will be computed with a unique set of attached molecular TR, uniquely assigned to every molecule. The numerical values of the computed QSM elements in all the cases tested have been shown to be inferior to the bimolecular optimal results, as expected.

To have such a sequence of QSM applying the Algorithm 2 to every cyclic permutation reordering, there is no alternative to considering a QSM for each molecule in the set:

\[
\forall l \in \{1, M\} : \exists Z_{il}
\]

The whole procedure can be easily parallelized, because the construction of every matrix \(Z_{il}\) can be performed independently from the rest. For every matrix in the above sequence, one can compute the complete sum of the off-diagonal elements:

\[
\forall l : S_l = (Z_{il}) - Tr(Z_{il})
\]

Among all the matrices generated, the optimal could be considered the one, which possesses the maximal complete sum of the off-diagonal elements. The MQSPS generated for each generated QSM and CSIM the sum of the matrix off-diagonal elements as a means of assessing the degree of optimality reached in the calculation.

A standalone F95 program has been constructed to apply Algorithm 2 for this circular permutation index. It is included in the MQSPS and described in the section Program Sequence Description below.

Some Considerations on Circular Permutation QSM

The circular permutation results can be manipulated in several ways, after the QSM set \(\{Z_{il} l = 1, M\}\) is known. One of the possible transformations consists of constructing a convex linear combination of the circular permutation QSM set elements. Given a convex set of scalars \(\{z_l l = 1, M\}\) : \(\sum_{l=1}^{M} z_l = 1\), then the convex linear combination:

\[
\Lambda = \sum_{l=1}^{M} z_l Z_{il}
\]

will act as an averaged QSM, representing the whole molecular set in a unique manner.

Random Index Permutations

A general index permutation algorithm and its application program can also be easily obtained through application of Algorithm 2. Even thinking of a parallel implementation of such an index permutation framework, one can simply foresee a formidable growing computational cost involving large molecular sets (\(M > 8\)), as the number of QSM computations will increase as \(M\). Therefore, to take into account, all permutations will be a task almost out of reach for molecular large sets. Even if run on modern computers: seeking for parallelization of every permutation choice of MS will become too heavy a load if all the permuted indices are tested.
However, there are several possible ways to circumvent this computationally explosive permutation problem. A simple way consists of choosing at random some permutations and searching for that which provides an optimal QSM, as in the circular permutation system described earlier. Another possibility could be the construction of some kind of Monte Carlo procedure, accepting the random permutations with the index configuration providing an optimal QSM trying to keep the number of random trials reasonably low, while constructing a convex average QSM, or choosing as representative the QSM with the larger complete sum of off-diagonal elements, as proposed in the circular permutation case. The MQPS contains three Monte Carlo codes, which perform such random optimal searching; they are discussed below.

The Measure of the Optimality of the QSM Algorithms

In all the previous procedures, where by means of TR two molecules are superposed, the measure of the optimal MS relies on the magnitude of the chosen QSI. In this article, the simple overlap QSI as a QS measure has been selected in order not to increase the large amount of information one can retrieve from the set of available programs, including other QS measures. Still, other measures included in the structure of the QSI (8) can be used for the same purpose and in the same manner as in the present program set. It is only necessary to substitute the overlap QSI subroutine by the chosen alternative QS measure.

Thus, it is obvious that two molecular structures are better superposed in a given TR frame the larger is the QSI computed between them. In fact, Algorithms 1 and 2 use this criterion to choose the appropriate optimal TR coordinates involved in QSI evaluation.

This can be measured even in a general manner, when observing the complete sum of the off-diagonal elements of the QSM or CSIM. Greater QSM complete sums preclude a better MS scheme for the connected molecular set as a whole. Of course, the complete sum of the matrix made up of the SqDM’s complete sums for each molecular pair will accordingly decrease. Some tests have been performed taking this property of the SqDM, but they proved that this property is not such a strong criterion as the QSI maximization. However, the computational simplicity involving the SqDM tests deserve further research in the future.

A QSM Optimization Criterion

Hence, when the task of computing all the elements of a given QSM is finished, within the same computational framework, one can compare two QSM and decide, according to some criterion, which one has been better optimized. This can be simply done computing the complete sum of the superior (or both) triangle(s) of the QSM or CSIM pair to be compared. It must be taken into account that QSM and CSIM are symmetric matrices. Given two or more QSM, differing only in the TR scheme used when their QSI elements have been evaluated, the optimal, accordingly, will be that yielding a larger value of its complete sum of off-diagonal elements and the same can be said for CSIM. Such a simple criterion has been put forward along this article and used to test several possibilities for preparing the initial molecular atomic coordinates. The value of the sum of the two QSM or CSIM triangles is shown as the current output information in some of the MQPS F95 codes attached to this study.

Other criteria can be surely designed, but in view of the large number already present in this article, we leave these for future work.

A Restricted Elementary Jacobi Rotation Refinement of QSM

One can envisage the design of other conceivable refinements, not involving MS or QSI computation to be applied on already computed QSM. Usually, once computed over the overlap QSI or any QSM with a positive definite weight operator, such a QSM possesses positive definite elements.\(^{11}\) Off-diagonal elements of QSM depend on the relative positions of the involved molecules and are functions of TR manipulations performed on the molecular coordinates, as discussed in preceding sections.

Diagonal QSM elements are invariant tough, when the same TR operations are performed on the molecular coordinates of the active molecule making up the QSS diagonal element. Logically, when dealing with diagonal QSM elements, the maximum MS appears when the involved structure possesses exactly the same atomic coordinates in the diagonal QSI, and thus such QSSI becomes maximal.\(^{11}\)

Therefore, when optimizing QSM elements, the diagonal ones are left invariant, while one seeks for maximal off-diagonal values via TR operations. This has been already discussed when describing Algorithms 1 and 2.

Consequently, optimal QSM operations can be seen as types of transformations performed on the QSM elements, resulting in maximal values in the off-diagonal elements, while leaving diagonal elements invariant.

To have a detailed view on the above problem, one can start supposing a \((2 \times 2)\) QSM like:

\[
Z = \begin{pmatrix}
\theta_1 & z \\
z & \theta_2
\end{pmatrix}
\]

where there is only one off-diagonal element \(z\) and two diagonal \(\{\theta_1;\theta_2\}\). Then, an EJR\(^{17}\) matrix can be written as follows:

\[
J = \begin{pmatrix}
c & -s \\
s & c
\end{pmatrix}
\]

where \(c\) and \(s\) stand for the EJR cosine and sine, respectively.

\(^{11}\)In some cases, the density gradient QSI, as it has been shown in the Part I\(^{10}\) of this series, can be negative, but this characteristic only appears when MS is far from optimal.

\(^{12}\)Evidently, when the QSS calculation is performed, but moving in a set of TR, one of the two involved identical structures, then the QSS measure is not optimal unless there is reached a perfect MS.
The transformation of the above defined $(2 \times 2)$ QSM via the EJR yields:

$$A = F^T Z J$$

$$= \begin{pmatrix}
    \theta_1 + (\theta_2 - \theta_1)x^2 - 2\varphi c & (\theta_1 + \theta_2)c + z(c^2 - s^2) \\
    (\theta_1 + \theta_2)c + z(c^2 - s^2) & \theta_2 + (\theta_1 - \theta_2)s^2 + 2\varphi c
\end{pmatrix}$$

The variation of the diagonal elements is essentially the same for both items, but with reversed sign. Thus, the constrained EJR transformation has to yield a constant variation in both diagonal terms, that is:

$$(\theta_2 - \theta_1)x^2 - 2\varphi c = 0 \rightarrow (\theta_2 - \theta_1)s = 2\varphi c \rightarrow t = \frac{2\varphi}{\theta_2 - \theta_1} = \frac{z}{\delta},$$

where $t$ represents the tangent of the EJR rotation angle.

On the other hand, because the double angle trigonometric cosine and sine of the EJR are $S = 2\varphi c$ and $C = c^2 - s^2$, the variation of the off-diagonal element can be simply rewritten as follows:

$$\omega(S) = \frac{\varphi}{\delta} S + zC.$$

Also, using $\varphi = \frac{1}{2}(\theta_1 + \theta_2)$ and $C = (1 - S^2)^\frac{1}{2}$, the $\omega(S)$ function results into a unique sine function:

$$\omega(S) = \frac{\varphi}{\delta} S + z(1 - S^2)^\frac{1}{2}.$$

A simple procedure can be designed supposing one is searching for the function $\omega(S)$ with a value as maximal as possible, but subjects to the zero diagonal EJR variation constraint. This provides the following relationship between sine and cosine:

$$s = \frac{z}{\delta}c.$$

Then, one can also write:

$$\omega(s) = \varphi \delta 2c + z(1 - 2s^2),$$

which can be rewritten in turn as a function of the EJR tangent and cosine:

$$\omega(t, c) = \varphi \delta tc^2 + z(1 - 2(tc)^2).$$

It is well known that the cosine can also be expressed in terms of the tangent as follows:

$$t = \frac{s}{c}, \quad c^2 = \frac{1 - t^2}{c^2} \rightarrow c^2 = (1 + t^2)^{-1}.$$

Therefore, one can finally write the off-diagonal function variation as a function of the tangent only, simplifying to give the expression:

$$\omega(t) = z + 2t(\varphi - zt)(1 + t^2)^{-1}.$$  

Using the tangent, which leaves invariant the diagonal elements, one can seek for some EJR, which maximizes the off-diagonal elements.

This can be performed using the molecular set pairs in a given QSM one at a time, accepting the resultant EJR $\omega(t)$ value in the corresponding off-diagonal element, whenever it is increased or leaving it invariant if it decreases. This maximal condition for the increment implies:

$$2t(\varphi - zt)(1 + t^2)^{-1} > 0.$$  

Therefore, choosing the EJR tangent always positive, that is with the convention:

$$\varphi = |\theta_2 - \theta_1|,$$

the above condition relies just on:

$$\varphi - zt > 0.$$  

Now using the constrained EJR tangent value $t = \frac{z}{\delta}$, one can also write:

$$\varphi - zt > 0 \rightarrow \frac{\varphi}{\delta} - \frac{z^2}{\delta} > 0.$$  

When this condition holds, then one can substitute the original QSM off-diagonal element value $z$ with the corresponding EJR augmented one. When this restricted EJR does not comply with the above condition, then, as already commented, the QSM off-diagonal value is left invariant. Moreover, although the computational practice, which accompanies this article, is based on atomic shell approximation (ASA) DFs, the GQSMS described algorithmic variants put forward in the present contribution can be applied within any theoretical level of QSI evaluation.

Contrarily to the Jacobi diagonalization procedure, the presented algorithm just applies in the $(2 \times 2)$ submatrices of any QSM, without modifying the rest of the off-diagonal elements entering the whole QSM itself.

In this way, the EJR-modified QSM can be used as a refined QSM structure without either further molecular alignment or QSI computation. Moreover, the described EJR procedure can be used on any kind of QSM irrespective of the operator used.

This restricted EJR algorithm has been implemented and tested by means of a standalone E95 program forming part of the MQSPS, see below in the section Program Sequence Description. As in some of the QSM obtained via molecular pair superposition, this EJR algorithm produces refined QSMs, which are not positive definite.

**Scaling and Inward Functions of QSM**

QSM scaling can be easily made, thinking of the simple transformation of the DF into a shape function: just dividing the original density by the number of electrons. Thus, for any QSI $Z_{AB}$ involving two molecular structures $A$ and $B$, possessing $\{\nu_A; \nu_B\}$ electrons, respectively, the corresponding QSI over shape functions can be written as follows:

$$\forall A, B : Z_{AB} = (\nu_A \nu_B)^{-\frac{1}{2}} Z_{AB},$$

which corresponds to a trivial matrix transformation of the QSM, quite similar to the one used when evaluating in a general manner the CSIM. Indeed, suppose a diagonal matrix such as:

$$\begin{pmatrix}
    d_{11} & 0 \\
    0 & d_{22}
\end{pmatrix}$$
\[ N = \text{Diag}(\gamma_i; i = 1, M) \rightarrow N^{-1} = \text{Diag}(\gamma_i^{-1}) \]

The shape scaled QSM (ShQSM): \( \Theta = \{0_{ij}\} \), which corresponds to the original QSM: \( Z = \{Z_{ij}\} \), is easily obtained by means of the matrix equation:

\[ \Theta = N^{-1}ZN^{-1}. \]

Such kind of ShQSM construction can have some effects on the properties of the QSM, besides producing matrix elements with numerical values, which appear smaller than in the original form. However, CSIMs in both original QSM and ShQSM appear to be the same; thus, certainly KMST and therefore KTO remain invariant upon this shape scaling.

Nevertheless, this is not the only possible scaling or transformation of the QSM elements which can be performed on the individual QSI. Other possibilities are useful and will be briefly described below.

**Inward Function of a Matrix**

Under the so-called inward operations (see refs. 23–27 for more information), matrices of the same dimension behave almost like scalars. Among other inward operations, the inward function of a matrix can be easily described as follows. Suppose an arbitrarily defined matrix: \( A = \{a_{ij}\} \) the inward function of \( A; f(A) \), is a matrix of the same dimension as \( A \), simply defined as:

\[ f(A) = \{f(a_{ij})\}. \]

**Properties of the Square Root of a QSM**

The structure of scalar products and their generalization have been discussed recently. Because of this previous suggestion, it was considered interesting here to compute the inward square root (ISQR) of a QSM and use the resultant matrix instead of the original QSM matrix.

The ISQR of a QSM has been used in several published articles, dealing with QS and quantum quantitative structure–property relationship (QQSPR) problems. Such a matrix, which can be written as \( Z^{\left[2\right]} \), can be easily defined, as just a trivial algorithm is needed, such as:

\[ Z = \{zu\} \rightarrow Z^{\left[2\right]} = \{\sqrt{zu}\}. \]

The ISQR of a matrix with positive definite elements, like the usual QSM, has the advantage that the physical dimensions of scalar products and norms, resulting from the product of both vectors, become the same as those of the involved vector components. This is a well-known and widely used property, for instance in statistics, where one can quote a typical example, which appears in the pair of statistical parameters: variance and standard deviation.

A specific standalone F95 program forming part of the MQSPS has been devised to handle such a QSM transformation, and it is described in the Program Sequence Description below.

The obtained results are interesting enough to deserve some discussion. Naturally, the transformed QSI becomes smaller than those contained in the original QSM. Nevertheless, it seems that the implicit ordering enclosed in the original KMST is preserved within the structure of the ISQR QSM. Thus, such a QSM transformation seems to contain the same type of information as in the original QSM, as in the shape scaling discussed above. The advantage of the ISQR consists of the appearance of smaller matrix elements in \( Z^{\left[2\right]} \) than in the original QSM \( Z \). Another characteristic of such ISQR QSM appears when the eigenvalues of such a matrix are computed. As usually, the QSM possesses all positive eigenvalues or very few negative ones; ISQR QSM has instead an augmented subset of negative eigenvalues.

Take as an example the Cramer steroids set, which will be discussed below in more detail in the computational examples section. The 21 steroid Cramer set has been handled with ASA densities computed with Mulliken atomic populations, obtained in turn at the HF computational level with a GTO basis set 3-21G using the Spartan code. The same calculation has been calculated in the ASA framework with Huzinaga basis functions and constructed with the QGMS algorithm, described previously in this study and applied in Algorithms 1 and 2. Under these circumstances, the Cramer set spectrum QSM is positive definite. However, the corresponding ISQR QSM possesses a negative eigenvalue.

**Other Inward Functions of QSM**

Alternative inward functions of QSM can also be implemented and studied as substitutes for the inward square root. For instance, one can choose the natural (or decimal) logarithm and the arc tangent, which are included in any FORTRAN language level as standard intrinsic functions. All these inward function QSM possibilities have been tested by means of a standalone program, including the previous shape scaling and square root, which was located into the MQSPS, associated with this article, and commented in the corresponding Program Sequence Description below.

**Program Sequence Description of the MQSPS**

Seven standalone programs forming the MQSPS have been constructed as a mean to illustrate this work from the computational point of view. The basic information structure is a F95 module, which serves to represent a defined type in any other program or subroutine, allowing a representation of any molecule in the ASA framework. Appendix C shows the structure of such a module. More details about the ASA computational setup can be obtained from the basic development papers 49–53, and modern additional information can be found in refs. 10 and 18.

Within the MQSPS codes, associated with this paper, there are two text files, which correspond to the needed flags and labels for each piece of code. The flag file (Flagstaff_n.txt) contains the indications needed for the options in the program, like verbosity in the follow-up output file, a priori atomic coordinate manipulation, etc. The file enters 0 and 1 flags, which are transformed within each program into a logical variable false-

\[ ^{18} \text{Where the parameter } n \text{ identifies the flag or label file and corresponds to the associated MQSPS code, which is using it.} \]
true content, respectively. The label file (Labelstaff_n.txt)\textsuperscript{44} handles the labels, which identify in a general manner the input and output files used in any of the codes of MQSPS. The label file permits the creation of input–output files prefixes and suffixes, besides choosing a generic file name for each molecular set.

The same flexibility cannot be associated in the MQSPS for the program array dimensions. The user has to modify them, if necessary, in each of the F95 source codes and recompile the modified program.

The programs, which are available to the readers via downloading from a web site,\textsuperscript{28} are briefly discussed here:

1. MQSS 1: constructs the data for a given molecular set, which is defined by means of minimal information input. The information consists of a coordinate file and, if the choice is not a promolecular ASA density, then an atomic population file is read. The output of this program can be read without further manipulation except the file labels, by the rest of programs in the MQSPS. In any case, the tag and label files have commentaries included where the meaning of every item into the file is explained. To simplify the program description, one can call IN1 and OUT1 the needed input and the elaborated output. The input labels permit to choose between three possible ASA basis sets: 3-21G, 6-311G, and Huzinaga. Other varied options for ASA basis sets can be obtained from ref. 57. These three chosen basis set options are present in the program folder as three files, which are called by means of a flag within the label input file. In addition, IN1 permits at will the use of Atomic Numbers or either Mulliken or Natural Orbital Atomic Populations. Atomic populations can be substituted by atomic charges too, if needed. With a convenient label, the Atomic Numbers are transformed into valence electrons, in case the users want to compute QSM with just valence promolecular densities. Also, an ASA bulk atomic density (ASA BAD) facility is also present, where the QSMs are calculated using a sum of the ASA contributions in a given molecule with unit atomic population coefficients. Thus, for every molecular batch a large number of possible choices are already present for generating the output file OUT1. Consider the three ASA basis sets available and the three atomic condensed densities that can be used. Combining them one can obtain nine variants for the same molecular set of atomic coordinates. The output file OUT1 can be read by all the other programs in the code series MQSS \( n = 2,7 \) of the MQSPS. The original atomic coordinates for each molecule can be made uniform by using appropriate flags. Thus, the coordinates can be translated to a unique center of atomic charges origin, the nuclear dipole moments can be reoriented lying along the \( x \)-axis, and the coordinate axis can be reoriented according to the principal components of the nuclear quadrupole moment.

2. MQSS 2: This program uses Algorithm 1 to construct the QSM of the molecular set represented by the file OUT1, which is used as input. The molecular coordinates are TR optimized for each molecular pair using overlap ASA similarity integrals, which are computed via a subroutine. It is straightforward to use other QSI within the program; however, at the present status of the codes, this adaptation must be carried out by the users. The program, among other output files, gives Z2, the QSM of the molecular set and C2, the corresponding CSIM. The optimized coordinates for every molecular pair are given in XYZ2.

3. MQSS 3: Reads Z2 and C2 to compute the KMST of the molecular set, which can be found in the output file KT3, and the KTO is provided in the output file KO3.

4. MQSS4a: Reads OUT1 and KO3 to provide as output, via Algorithm 2, the matrices Z4 and C4, the new QSM and CSIM computed with a unique optimal set of coordinates within a FAQS procedure, which are saved as the output file XYZ4.

5. MQSS 4b: This program is similar to the MQSS 4a, but performs a type of Monte Carlo search of a limited number of random index permutations. It produces the matrices Z4b and C4b and XYZ4b as output.

6. MQSS 5: This program reads OUT1 to construct via cyclic, or alternatively with a subset of random, canonical index permutations the QSM and CSIM files for each index permutation using Algorithm 2. The output files Z5, C5, and XYZ5 can be used for further calculations or manipulations. The result is a FAQS set of matrices.

7. MQSS 6: Reads any of the previous Molecular Similarity programs QSM files and refines them by using the restricted EJR algorithm. It returns the refined QSM: Z6 and the corresponding CSIM: C6.

8. MQSS 7: Using any of the QSM files, generated by the preceding Molecular type programs, computes (given the label file containing any of the extra prefix labels: XAPE, SQRT, LOG, or ATAN) the corresponding Z7 and C7 matrices, associated to either QSM shape scaling or the three Inward Matrix Functions. In addition, the KMST and KTO for any chosen Inward Function are given as output files.

Final Considerations

As a whole, the proposed standalone MQSS \( n \) \( (n = 1,7) \) MQSPS provides abundant output as already mentioned. Such output can be even enlarged whenever in the MQSPS codes MQSS 2, 4, and 5, other QSI are programmed, instead of the already implemented overlap QS measure. To keep the MQSPS code as simple as possible, only an overlap-like QSI subroutine is currently available, as already commented. Implementing other QS measures is straightforward.

The output of the MQSPS codes can be used as input in other elements of the series. For instance, the output of MQSS 4a and 4b programs can be used as input into MQSS 3, via the label file, after changing the names of the text files associated with the generated QSM. This means that, whenever a code belonging to the MQSPS, issues some output file containing a QSM, this file, conveniently renamed via the appropriate label file, can be used in any other program within MQSPS having any QSM as input.

Thus, some iterative structure not explicitly programed can be performed in this way. For instance, the KTO of an optimal bimolecular QSM (computed using MQSS 2), obtained as output of MQSS 3, can be used to produce a frozen Aufbau QSM, using the MQSS 4a program. The frozen Aufbau QSM obtained
in this manner can be used, in turn, to obtain a new KTO using MQSS 3, and so on.

ASA alternative atomic data can be easily downloaded from the web site given in ref. 57. The source codes discussed in this work are presented in F95 language and can be found in the web site of ref. 28, where input–output files are described and some examples are also given. Under the Windows 7 operating system, program compilation has been tested using two F95 compilers, provided by the web sites of refs. 58 and 59.

The MQPS codes have been tested on the examples described below. The executable files were on desktop and portable PCs (with at least 2GB of RAM). The executable and input–output files needed in or generated by the MQPS have been transferred to and executed on USB portable disks.

A final remark is due, as the codes of the MQPS have been tested with two kinds of F95 compilers, see refs. 58 and 59. In the programs involving random calculations, the random number generator routines differ if different compilers are used. However, the users can find out easily these lines within the programs and change them if the compiler is incompatible with the ones used here.

Assorted Application Examples

Introduction

Several molecular sets have been chosen to test the MQPS. The entire input and output files for every program are included in the corresponding subfolders within each program folder. All input and output files are text files (with extension .txt) in order that readers can readily access this information.

Geometries and Mulliken gross atomic populations for all the molecular structures have been obtained at the Hartree-Fock theory level by using a 3-21G basis set. The Spartan program55 has been the source of the data needed before performing QS computations.

It has earlier been commented a number of times that the possible amount of generated input–output files can be large and overwhelming. For this reason, the options, chosen to present some output results, are limited to keep the printed information reasonably concise.

Description of the Sample Molecular Sets Used

The MQPS has been used with the following molecular sets:

- Eight varied molecules: {(1) 2 methyl propane, (2) 1,1’difluoro,2-fluoro-ethane, (3) estradiol, (4) glycine, (5) benzene, (6) borazine, (7) ascorbic acid (vitamin C), and (8) zidovudine (AZT)}.

This was intended to represent a collection of quite diverse structures. Estradiol is an element of the Cramer 21 steroid set described in the following paragraphs and referred previously. Glycine, benzene, ascorbic acid, and zidovudine are extracted from the Spartan55 built in files and recalculated at the aforementioned theory level.

- Curious molecules 10: [atracine, calcitrol (vitamine D), piccine, picrocidin A1, ritalin, streptochlorin, thalidomide, estrone, 17a-ethinylestradiol, and levonorgestrel].

This set is made up of a collection of 10 structures, which except for estrone, which is an element of the 21 Cramer steroid set described below, has been taken from the recent newsread commentaries appearing in the ACS bulletin C&E News, along with the issues of March, 2010. The molecules are of relevance in areas ranging from the technological to the biological. The set contains one of the largest structures enclosed in this batch of examples: calcitrol (72 atoms).

- F- and Cl-methanes: (1) CH₄, (2) FCH₃, . . . (21) Cl₃C.

This sample set is constituted by the fluorine and chlorine derivatives of methane. These molecules have been chosen because of the simplicity of the structures compared with the other examples in this series.

- 21 Cramer steroids29: {(1) aldosterone, . . . (21) testosterone}.

This sample set has been chosen because the 21 Cramer steroids are large molecular structures (about 50 atoms), which have been used in many QS applications before31–34 the present one. For this reason, precise details of the Cramer set are not given here, as they are sufficiently described in the literature. The original order of the molecules has been preserved from the literature studies, so the reader can have a look at any of the quoted papers to obtain molecular names, structure drawings, and ordering indices.

- 12 Flavonoids60: Flavon derivative are ubiquitous molecular structures present in plants. They have a wide range of biological effects, apparently all of them with pleasant effects in diet, health, and curative activity in many fields. See also, for example, refs. 60–65. The 12 structures chosen here have been obtained from ref. 60. The ordered structures are as follows:

  1. (1) I Orobol [0.670]; 2. (2) I Formononetin [−0.230]; 3. (5) I Biochanin A [−0.090]; 4. (6) II Luteolin [0.640]; 5. (7) II Chrysoeriol [0.110]; 6. (10) II Hipsidulin [0.040]; 7. (11) II Galangin [0.860]; 8. (12) II Quercetin [0.780]; 9. (14) II Morin [0.810]; 10. (15) II Myricetin [0.830]; 11. (17) II Isorhamnetin [0.810]; 12. (18) II Limocitrin [0.700].

The first number in italics is the present ordering, the second in brackets is the ordering in ref. 60, the Roman figures correspond to Type I and II flavonoids, as given in the mentioned reference, the name of the structure precedes the biological activity (taken from ref. 65) in square brackets. The reported biological activity of these compounds corresponds to lipid peroxidation inhibition potency60 (%ILPO).

The molecular sets reported above have been used separately in all the described MQPS codes. The QS calculations along the MQPS have been done using the ASA in a molecular formulation, where the atomic density weights are chosen as Mulliken gross atomic populations, see Parts 1 and 2 of these commentaries on QS for more information.10,18

The ASA basis set used has been the Huzinaga, as contained in the file Huzinaga.txt of the program folder: MQSS 1. Such an ASA basis set choice with also the Mulliken gross atomic populations as atomic ASA weights has been made to present a set of results as homogenous as possible. The QS theoretical level used in any case has been only the density overlap QSI measure.
Every program folder in the MQSPS contains both input and output file subfolders, where sub-subfolders hold the input–output information on the five considered molecular sets.

On the Positive Definite Nature of the QSM

The sample molecular sets described above have been used to evaluate several QSM. By means of the program contained in the folder MQSS 2, the bimolecular atomic coordinates TR, associated to Algorithm 1, has been performed on the chosen sets to evaluate the bimolecular QSI. Based on overlap density QSI, the respective QSMs are positive definite with one exception: the 12 Flavonoids molecular set. This can be checked in the output subfolders of the folder MQSS 3. This feature is a clear example of the MS correct behavior under the proposed GQSMS algorithm. However, the 12 Flavonoids behavior demonstrates that there cannot be a proof that all bimolecular optimized QSM will be positive definite.

In fact, as an alternative test, a set of 21 molecules with four and five nonhydrogen atoms has also been studied. These are not included in the present web batch. When the associated QSM is computed within the same background and ASA frame as the five reported molecular sets, it yields, within the bimolecular optimization of MQSS 2 code, a spectrum bearing also one negative eigenvalue. However, the resulting QSM performed in a FAQS KTO computation framework yields indeed a positive definite QSM. This result does not prove that, in general, the FAQS procedure provides positive definite QSM, but at least it furnishes an outcome in the appropriate direction. In fact, such a result is theoretically expected, as the ensuing QSM, although of less Q quality than the bimolecular ones, in principle corresponds to metric matrices, associated with a sole set of molecular TR DFs, attached uniquely to every molecule and consequently they are positive definite matrices.

As a consequence of having recalculated the QSM, but with the scheme of the FAQS algorithm or the Monte Carlo technique, which give similar results, the subsequent unified coordinates produce matrices, for example, for the 21 Cramer steroid set, which continue to be positive definite.

The same cannot be said when the EJR refinement or the inward functions are evaluated over the original bimolecular TR QSM. These transformed QSMs possess from one to three negative eigenvalues. Such a result can be attributed to the fact that the resultant modified QSMs are no longer representative of a metric matrix as in some of the previous cases. Even with this nondefinite spectrum characteristic, the modified QSM can be alternative candidates for molecular set ordering and can furnish distinct parameters in QQSAR applications.

Kruskal Tree Representations

KMST can be easily drawn from CSIM computed from the appropriate QSM, as one can see from the attached figures associated with the proposed molecular sets. The examples presented are sufficient to grasp the power of QS to order in many ways molecular sets of any kind. Such a possibility provides a naive example of how the Axiom of Choice, see, for example, ref. 66, can be applied to any collection of molecules. Upon accepting the quantum mechanical representation of the molecular structures in a given set, one can order them in some kind of molecular periodic table. One way to obtain this ordering is by using KMST and the derived KTO. Such a general possibility has been put forward several years ago by two of us.67

Figures 1–5 depict the KMST for the molecular sets, described at the beginning of this section. For the first two molecular sets, Figures 1 and 2 show what one expects when two sets of almost the same cardinality, but quite different elements, are ordered according to the KMST algorithm.

The results can be summarized in the following figures, which are briefly discussed.

With respect to the molecular set attached to Figure 1, the bimolecular similarity analysis yields molecules 5 and 6 as the most similar from the density overlap point of view. This corresponds to the benzene–borazine pair. Borazine transforms afterward into a crossing point where molecules 1, 2, and 4 met. These three structures correspond to 2-methylpropane, 1,1′-difluoro-2-fluoro-ethane, estradiol, glycine, benzene, borazine, ascorbic acid (vitamin C), and zidovudine (AZT).
The molecular set curious molecules 10 also corresponds to a quite heterogeneous collection, even more so than the previous set. As it has been commented, the structures within this set have been collected from the March issues of C&E News, except for estrone, an element of the 21 Cramer steroid set. This particular molecular set choice has been made in order to do not select molecular structures forming a set, which can be biased by the will of the authors, but having a certainly independent character as much as possible. The set has been essentially chosen because it illustrates again the power of the proposed MS algorithm to permit the ordering of molecular sets independently of their nature. According to the bimolecular overlap QSI, atracine (1) and streptochlorin (6) are the most similar structures of the collection. This is essentially because of the presence of a chlorine atom in both molecular frames; superposition based on overlap similarity backs this fact up. Besides this connection, which structures the KMST backbone, there is no other relevant feature, as the molecular structures gathered here are so diverse, except that the poly-benzenoid hydrocarbon: (1) picene, appears as a marginal molecular structure at the end of one of the tree branches. In addition, the three steroid-like molecules: (8) estrone, (9) 17a-ethinylestradiol, and (10) levonorgestrel also appear in some marginal positions of the tree.

The ordering of the F- and Cl-methanes set provides another interesting picture associated with a Kruskal tree. In this case, the KMST is star shaped with structure (6), ClCH₃, at the star vertex. The larger bimolecular overlap similarities are those between this structure and the molecules (10): (FClCH₂), (11): (F₂ClCH₂), and (12): (F₃ClCH₂). This is due to the superposition
of the Cl atom in the four structures before considering any other substructure. Moreover, methane (1) is also connected to CICH₃. One expects this kind of information within this sort of molecular sets. An appropriate KTO has been found anyway, despite the heavy-weight Cl atoms play in the MS overlap QS scheme.

In the KMST of Figure 4, the most similar structures are (11): (11-deoxycortisol) and (20): (17α-hydroxyprogesterone). One can easily realize that the main difference between both molecules appears in the substituent endings: (11) \(\sim (\text{OH})\) and (20) \(\sim (\text{CH}_3)\) at the terminal position 17; the rest of the molecular skeleton is coincident.

Although there are other coincident structural pairs in the 21 Cramer steroids collection, the GQSMS algorithm in conjunction with overlap QSI has chosen this molecular pair to start the KMST sequence. Next to it appears the structure (10) (11-deoxycorticosterone) and connected to this molecule (6) (corticosterone). The pattern corresponds to three structures: (6), (10), and (11), which only differ in the displacement, presence, or absence of an hydroxyl moiety. These structures represent the left branch of the 21 Cramer steroids set KMST, which somehow ends with the structure (1): (aldosterone) connected to structure (6). Such ending place is reasonable as aldosterone is structurally far away from any molecule in the 21 Cramer steroids set. From the overlap QSI point of view, however, structure (1) is the nearest possible to (6).

An interesting final comment, which makes evident a qualitative attempt at a structure–property relationship, present in the KMST, can be associated with the binding affinity of the steroid set as given originally in the ref. 29, which is also sketched in ref. 35. Indeed, the main KMST structure made of the four steroids, (11), (20), (19), and (10), corresponds to testosterone-binding globulin affinities measuring \([-7.2]\), \([-6.9]\), \([-6.9]\), and \([-7.4]\), respectively, constructing thus a certain activity island trend. Such activity trends among the KMST branches are repeated again in the sequence: (6), (7), and (8), corresponding to corticosterone, cortisol, and cortisone, associated with quite coincident activities \([-6.3]\), \([-6.2]\), and \([-6.4]\), respectively. Aldosterone (1), already commented as being an isolated structure, connects with (6) and possesses an activity of \([-5.3]\).

Other comments can be made on Figure 4, but they are left to the readers as other substructures one can find in the KTO can be obtained by means of similar straightforward reasoning as provided before.

The 12 flavonoids set studied here is composed of three Type I: \{(1), (2), (3)\} and Type II: \{the rest (4)-(12)\} flavone derivatives. Type III derivatives included in the original reference\(^60\) have been discarded here because of the quite different structural form Type III flavonoids possess, compared with Types I and II. The used molecular structures and names have been given beforehand and have been collected from ref. 60. Other Type I molecular structures have been ignored too because of the large glycoside substituents they possess.

The obvious KMST picture appears, when one observes in Figure 6, that the Type I molecules correspond to the branch ends of the KMST structure, whereas the central part of the tree corresponds to the Type II structures. Moreover, in the KMST of Figure 5, one can see how molecules \{(8), (9), (10), (11), (12)\}, corresponding to the subset of greater %ILPO, are clustered together.

In addition, some more remarks on other issues of the MQSPS codes can also be made. In the MQSPS available into the web, there are ways other than the classical to obtain KMST and thus KTO. For KMST issued from the inward functions in MQSS7, it seems that systematically one of the molecules in all the sets tested appears as the center of a star shaped KMST. The reader can observe this detail in the output subfolders of the folder MQSS7.

Moreover, it has been observed, alongside the multiple tests performed before deciding on the examples to be shown, that KMST are in their central part quite stable, with the larger similarities also stable along the chosen coordinate preparation and TR modes. Thus, the KMST figures presented here have their main features stabilized in such a way as to be trusted to correspond to some inner set ordering structure.

**Molecular Superposition Visualization**

The 21 Cramer steroids and 12 flavonoids molecular sets have been chosen in addition because it is possible to visualize the
result of the MS of the whole set quite easily. Both molecular sets are constructed with a common skeleton, possessing in this way some uniformity along the members of the set and are certainly much less heterogeneous than the eight varied molecules or the curious molecules 10 sets.

The authors do not completely share the widespread belief that an image is worth thousand words. Although it is undeniable that an image can illustrate fairly well a result attached to a computational technique, in general, it hardly proves anything else, a theorem for instance, except in a few assorted cases. However, because of the present situation of the computational set of examples, resulting from the application of the MQPS code, then at this point of the development the authors feel that the results deserve to be illustrated. In this way, some visual application outcomes of the proposed GQSMS algorithm can be presented for these more homogeneous molecular sets.

Figures 6 and 7 have been both chosen, as the MS pictures they offer appear to be the most consistent with chemical intuition of all the obtained results. Heterogeneous molecular sets MS diagrams, as those of KMST in Figures 1 and 2, have much less interpretative power.

Final Discussion and Remarks

A novel way to compute QSM with a new MS algorithm, based on the properties of the SqDM, calculated between two molecular Cartesian frames has been described. The bimolecular GQSMS algorithm, issued from this study, can be easily extended to evaluate multiple densities QSM and it is easily parallelizable. Moreover, although the computational practice, which accompanies this article, is based upon ASA DFs, the GQSMS algorithmic variants put forward in the present contribution can be applied within any theoretical level of QSI evaluation.

Several QSM construction alternatives, based upon the QSAP, can be effortlessly described. The computation of KMST from CSIM derived from any QSM permits the construction of a reordered molecular index sequence: the KTO. Such a sequence can be used to construct new QSM under the algorithms associated to FAQS with a uniquely defined set of optimally oriented coordinates. Similar operations can be done with cyclic or random permutations of the canonical molecular indices.

The invariance of the diagonal elements of any QSM on optimization of the upper and lower matrix triangle elements, via TR of molecular coordinates, has inspired a restricted EJR procedure, which augments, if possible, these off-diagonal elements of any QSM, without need to compute neither TR of the molecular coordinates nor the corresponding QSI.

Finally, simple transformations based upon the use of shape functions to compute the corresponding QSI or the use of inward matrix functions over already known QSM permit an extension to the QS techniques used until now.

The MQPS, a set of F95 code implementations of all the ideas expressed in this article, are available to the readers and can be downloaded from a public web site.\(^\text{28}\)

All the computational tasks, program developing, and compiling, preliminary data on molecular structures retrieving, and subsequent QSM calculation and manipulation have been performed in portable and desktop computers. Migration to larger computational facilities and parallelization is a trivial technical matter.

Acknowledgments

L. D. Mercado work is associated to a research fellowship attached to this project. The authors thank the friendly work of Professor Neil Allan who has read and corrected the initial version of this manuscript.

Appendix A: Kruskal’s Minimum Spanning Tree Algorithm

FORTRAN 95

Subroutine Kruskal’s Tree(M,mm,Order,S)
!
integer Order(mm,3),S(M,2)
Logical TK, TL
!
!M is the number of molecules and mm=(M*M+M)/2
!
!(Order(IJ,1);Order(IJ,2);IJ=1,mm) contains the
(IJ) pairs of elements in a distance-like matrix ordered from shortest to largest distances or from a similarity index matrix from largest to lesser values.
!
!(S(P,1);S(P,2))(P=1,M-1) contains at the end the Kruskal’s tree links.
!
do I=1,M; S(I,1)=0; S(I,2)=0; End do
!
IZI=1
I=Order(2,1); S(1,1)=I; J=Order(2,2); S(1,2)=J
!
do I=1,mm
K=Order(I,1);L=Order(I,2)
!
!This is here because of existing the diagonal (IJ) pairs
!
if(K=.eq.L)Cycle
!
TK=.False. !‘K is not into the tree
do KL=1,IZI
if(S(KL,1)=.eq.K.OR.S(KL,2)=.eq.K)TK=.True. !‘K is into the tree
end do
TL=.False. !‘L is not into the tree
do KL=1,IZI
if(S(KL,1)=.eq.L.OR.S(KL,2)=.eq.L)TL=.True. !‘L is into the tree
end do
!
!To admit the pair (K,L) as two vertices into Kruskal
!tree, one of both molecules has to be already into the
!tree vertex list and the other has to be not yet into
!the tree vertex list.
!
if((TK.and..not.TL).or.(.not.TK.and.TL))then
IZI=IZI+1
Subroutine KTO(M,ISeqS,Kruskal)
!
! Use Kruskal Minimum Spanning Tree index pairs
! to construct a new molecular index ordered sequence
!
! M is the number of molecules
! ISeqS(M,2) contains the pairs of original order indices
! forming the Kruskal Minimum Spanning Tree
! Kruskal(M) contains the old indices
! with the new Kruskal Tree Ordering
!
Dimension ISeqS(M,2),Kruskal(M)
Logical TaK,TaL

Do i = 1,M; Kruskal(i) = 0; end do
!
Klux = 0
Do i = 1,M-1
K = ISeqS(i,1); L = ISeqS(i,2)
TaK = .False.; TaL = .False.
!
Do J = 1,Klux
If(Kruskal(J) = = K) TaK = .True.
If(Kruskal(J) = = L) TaL = .True.
end do
!
If(not.TaK) then; Klux = Klux + 1; Kruskal(Klux) = K; end if
If(not.TaL) then; Klux = Klux + 1; Kruskal(Klux) = L; end if
If(Klux = = M) exit
end do
!
End

Appendix C: ASA Molecule Type Module

module Set_Molecule!
parameter(natMX=70) !Maximum atoms per Molecule
parameter(maxfn=10) !Maximum ASA functions per atom
!
type Molecule ! ASA prepared information for a molecule
!
CHARACTER (LEN=80):: Identity !molecular structure identifier label
!
end type Molecule
!
end module ! Set_Molecule

References