Quantum molecular similarity measures (QMSM) as a natural way leading towards a theoretical foundation of quantitative structure-properties relationships (QSPR) *

R. Carbó, E. Besalú, Ll. Amat and X. Fradera

Institute of Computational Chemistry, University of Girona, 17071 Girona, Catalonia, Spain

Quantum molecular similarity measures (QMSM) are succinctly described and justified as a tool to obtain ordered patterns within a given set of molecular electronic structures. The nature of QMSM appears also to establish the way leading towards a discrete representation of a given electronic structure, when using a quantum mechanical framework, in the form of some n-dimensional column vector. As a consequence, quantitative structure-properties relationships (QSPR) can be considered, in general, to be coincident with a procedure to obtain the discrete approximate representation vector elements of some unknown operator whose expectation values can be associated with a chosen observed experimental property value.

1. Introduction

In the last fifteen years the theoretical and practical formalism of QMSM has been developed [1] in our laboratory and by other authors [2]. Much older, however, appears to be the idea to obtain empirical relationships between handy parameters and molecular properties [3], and recent procedures seem to be very successful as a tool to predict new molecular structures with taylor-made properties [4]. As recently QMSM have been used as parameters in quantitative structure-activity relationships (QSAR) [5], it seems the time has arrived to search for the possible practical formalism allowing QMSM to be used in QSPR or QSAR environments.

Beyond this initial landscape, the success of QSAR in the realm of molecular design seems to be beyond doubt. Also certain is the fact that no comprehensive justification, other than the empirical evidence and pragmatism, has so far been given to make fundament this situation. The continued successful use of QSPR techniques cannot be an aleatory product based on statistical factors only: it seems to

---

* A contribution of the Grup de Química Teorica de Catalunya.

preclude the evidence of an existing solid theoretical reason not yet described. A new idea, associated both with the QMSM theoretical framework and the quantum mechanical operator expectation value concept, will provide in the following pages a solid ground for this task.

For this purpose, in the present paper, we first give a short review of the definition and meaning of QMSM. The way leading to the discrete n-dimensional representation of a molecular set is presented afterwards. Next, the use one can make of the previous molecular discrete representation is discussed, introducing Mendeleev's postulates. Then, a brief overview of the QSPR techniques precedes the definition of the discrete expectation value concept and, finally, the theoretical foundation of QSPR is exposed.

2. QMSM

The basic idea underlying the concept of QMSM is trivial. Given two molecules, $\{m_A, m_B\}$, suppose the Schrödinger equation is solved at an arbitrary level for both structures, so the respective wavefunctions, $\{\Psi_A, \Psi_B\}$, for a given state of both electronic structures are known. A density matrix, $\{\rho_A, \rho_B\}$, connected with the respective wavefunction pair can be computed in the usual way [6].

Using a definite positive operator Ω as a weight, a QMSM involving the molecules $\{m_A, m_B\}$ is defined as the following integral:

$$z_{AB}[\Omega] = \int \int \rho_A(r_1)\Omega(r_1, r_2)\rho_B(r_2) \, dr_1 \, dr_2 , \tag{1}$$

where $\{r_1, r_2\}$ are sets of electron coordinates associated with the corresponding density functions. In this framework, the QMSM z_{AB} are non-negative real numbers. Originally [1a], the weighting operator was chosen as a Dirac delta function, $\Omega = \delta(r_1 - r_2)$, and the involved densities, $\{\rho_A, \rho_B\}$, as the first order density functions; when using this, the QMSM as defined in eq. (1) becomes the so-called overlap-like measure:

$$z_{AB} = \int \rho_A(r)\rho_B(r) \, dr . \tag{2}$$

Many other QMSM can be defined, even within a more general conceptual context, see for example refs. [1l] and [1n] for more details. Among all the various possibilities, the one most conspicuously used is the so-called Coulomb-like measure, defined as

$$z_{AB}[r_{12}^{-1}] = \int \int \rho_A(r_1)|r_1 - r_2|^{-1}\rho_B(r_2) \, dr_1 \, dr_2 , \tag{3}$$

which transforms into the Coulomb molecular energy when considering the self-similarity measure $Z_{AA}[r_{12}^{-1}]$. Also, triple [1k] or multiple [1l] QMSM can be defined

without other problems than those of augmenting the complexity within the defini-
tion of the involved integral measures. Triple QMSM are easily constructed when
considering a third molecule $\{m_C\}$ besides the initial set $\{m_A, m_B\}$. Then, using the
attached density $\{\rho_C(r)\}$ instead of the operator Ω in definition (1), the following
measure appears:

$$z_{AB;C} = Z_{AB}[\rho_C] = \int \rho_A(r)\rho_C(r)\rho_B(r) \, dr, \tag{4}$$

giving one of the five possible definitions of QMSM involving three density func-
tions. As another example, if Ω is substituted by the off-diagonal element of the
density matrix $\rho_C(r_1, r_2)$, an alternative form, different from the previous QMSM
definition (4), is obtained as the following integral:

$$z'_{AB;C} = \int \int \rho_A(r_1)\rho_C(r_1, r_2)\rho_B(r_2) \, dr_1 \, dr_2. \tag{5}$$

In any case, the previous discussion shows that a wide collection of QMSM can
be defined in a unique way when the molecular density matrices are known.

3. Discrete representation of a molecular set

Let us suppose a given set of molecular structures, $M = \{m_I\}$. To every element
of M can be attached a density matrix element belonging to a set $P = \{\rho_I\}$. Then,

$$\forall m_I \in M \rightarrow \exists \rho_I \in P \Rightarrow m_I \Leftrightarrow \rho_I. \tag{6}$$

Once a particular QMSM form is chosen, for every ordered pair of $P \otimes P$ a
matrix element can be computed:

$$z_{JI} = \langle \rho_J | \Omega | \rho_I \rangle \tag{7}$$

and so a *similarity matrix* $Z = \{z_{JI}\}$ can be build up. The computation of each inte-
gral in eq. (7) needs a process of optimization, in order to align the involved mole-
cules to obtain a maximal value of z_{JI}. This optimization process may transform
the natural positive definitness of the matrix Z because of the different relative
position in space for the same molecule in front of the rest.

Partitioning the similarity matrix Z into a row hypervector with column ele-
ments $\{z_I\}$,

$$Z = (z_1, z_2, \cdots, z_I, \cdots), \tag{8}$$

one can see that the column elements are

$$z_I = \{z_{JI}; \forall J\} \in Z \tag{9}$$

and z_I may be considered to be a discrete matrix representation of ρ_I within the set P acting as a basis, set. Thus,

$$m_I \Leftrightarrow \rho_I \Leftrightarrow z_I . \tag{10}$$

Furthermore, nothing opposes the use of auxiliary molecular sets $A = \{a_K\}$ and their attached densities $D = \{\gamma_K\}$ to augment the vector dimensionality, representing a given molecule on M. That is, a new vector can easily be constructed as the direct sum,

$$u_I = z_I \oplus v_I \in U , \tag{11}$$

where

$$v_I = \{v_{KI} = \langle \gamma_K | \Omega | \rho_I \rangle; \forall K\} \in V . \tag{12}$$

The final conclusion at this stage may be that by choosing a QMSM and after computing a similarity matrix, a vector representation of the molecular density function in a given molecular set is obtained. Thus, a vector representation of a given molecule in the chosen basis set of molecular density functions can be constructed as

$$\forall m_I \in M \rightarrow \exists u_I \in U \Rightarrow m_I \Leftrightarrow u_I . \tag{13}$$

Proceeding in this way, which can be extended to QSPR methods, one can say that to a given molecular structure m_I a *point-molecule* u_I is associated. A given collection $U = Z \oplus V$ of *point-molecules* will be called a *molecular point-cloud*.

4. Mendeleev's postulates, molecular set order and visualization

The molecular point-cloud $U = \{u_I\}$ as defined in eqs. (11) and (12) may be manipulated afterwards in order to extract information from its elements or to obtain new values which, in turn, can be used by other algorithms. Visualization of the set U may be very helpful as a tool to gather information on the relationships between members of the set M [11,1n]. This possibility has been used in various ways, as well as the related option to employ QMSM or derived similarity indices, obtained from the manipulation of QMSM matrix elements, to obtain some ordering among the elements of the set M. Any system which permits to be studied by means of quantum mechanics will be called a *quantum object*. The principles governing these possibilities have been called Mendeleev's postulates [11,1n]. They can be summarized as follows:

1. *Every quantum object* in a given state can be described by its *density matrix elements*.

2. *Quantum objects* can be compared by means of a *quantum similarity measure*.

3. Projection of a *quantum object set* into some *n*-dimensional space is always feasible.

4. A *quantum object set ordering* exists.

The Mendeleev's postulates can be associated with the following points of the theory: Postulate 1 is a usual quantum mechanical assumption, postulate 2 describes the starting point for the use of QMSM allowing the definitions of section 2, postulate 3 describes the reasoning carried out in section 3, and postulate 4 is nothing more than the application of Zermelo's theorem to the developed QMSM theoretical context. More details can be found in refs. [1j,1l,1m].

Postulates 3 and 4 permit a pictorial visualization of the set M, using the representation form of every molecule in the set M contained in the molecular point-cloud U, as described in section 3. Reference [1g] established the basic concepts underlying these procedures.

5. QSPR

Having reached this point that so far described the discrete representation of molecular structures and their possible use, one can realize that this fact is also connecting the previous formalism with parent theoretical procedures used to obtain information on QSPR or particularly on QSAR.

A typical QSPR procedure consists of attaching to every element $m_I \in M$ a vector $q_I \in Q$, whose elements are chosen in an empirical way from various sources. Some are certainly chosen as molecular atomic charges or quantum chemical related parameters, but others come from empirical sources like octanolwater partition coefficients or may even constitute a purely binary information variable; others, finally, bear empirical structural intuitive bonding schemes like the connectivity related indices [7].

However, the fact is that, although in quite different ways, QMSM and QSPR techniques both attach a vector to every element of M. As in the QMSM case, one can call this vector a point-molecule. Next step in the QSPR framework consists of connecting a given molecular property value π with the molecular vector representation q throughout a linear equation, such as

$$x^T q = \pi, \tag{14}$$

which can also be observed as a linear functional transformation of the discrete point-molecule q by means of a dual space vector x^T, a vector whose set of coefficient elements can be easily obtained using a standard least-squares calculation. In QSPR, unless one chooses in a very restricted way the elements of the point-molecule q, as discussed some years ago [1b,8], no direct meaning whatsoever can be attached to the elements of the vector x.

6. Discrete expectation values

The form of eq. (14) in a QMSM environment may be written in a parallel manner as

$$w^T u = \pi,\tag{15}$$

where the constant π role as a molecular property is preserved here too. However, contrary to eq. (14), to the coefficient vector w elements, which may be obtained by a least-squares technique as in the QSPR framework, one can attach a coherent meaning to the whole QMSM theory developed so far.

To prove this, let us consider again the *point-molecules* $u_I \in U$, which, as defined in section 3, are nothing but a discrete representation of the densities $\rho_I \in P$. The representation of the *molecular point-cloud* $\{u_I\}$ vectors is obtained in the space where the basis $P \oplus D$ is active.

Thus, $u_I \Leftrightarrow \rho_I, \forall I$. At the same time, as it has been employed when defining triple QMSM in eqs. (4) or (5), the density ρ_I has also the structure of a positive definite operator, which in the QMSM context possesses in turn the matrix representation of the *point-molecule* u_I. From the quantum mechanical point of view, given any observable O, a given Hermitian operator Ω attached to it must exist, so the expectation value $\langle \Omega \rangle_I$ in the system described by ρ_I may formally be obtained as

$$\langle \Omega \rangle_I = \langle \Omega | \rho_I \rangle = \int \Omega \rho_I \, d\mathbf{r}.\tag{16}$$

Then, to the operator Ω there can be attached the discrete vector representation w using the same basis set as contained in $P \oplus D$, in such a way that both vectors u_I and w belong to the same discrete n-dimensional space representation. Thus, with the scalar product

$$\langle \Omega \rangle_I \sim w^T u_I\tag{17}$$

can be associated the approximated expectation value computed within this space.

The contents of this section and the related material from section 3 are a consequence of the usual computational practice in quantum chemistry and related quantum mechanical applications. Although they may appear unfamiliar to a reader, accustomed to square matrix representations of operators, it must be kept in mind that square matrix vector spaces may be made isomorphic to column matrix vector spaces of the appropriate dimension. A very good exposition of the fundament of all that is said here, although in a somewhat different context, mainly attributable to the different discussed applications, can be found in the monograph of Bohm and Gadella [9].

7. Theoretical foundation of QSPR

As it has been discussed earlier, every molecular property can be seen as some expectation value of some unknown operator whose matrix representation elements may be evaluated by means of eq. (17) using a least-squares technique. A more general form of eq. (17) may be considered here. Let us define a new vector of QMSM origin obtained by some, even non-linear, transformation of the original similarity vector space:

$$g = R(u),\qquad\qquad(18)$$

where $R(u)$ represents any possible mathematical manipulation of the point-molecule u elements; then the equation

$$w^T g = \pi\qquad\qquad(19)$$

constitutes a QSPR-like equation, deduced from purely QMSM theoretical considerations. There is thus a capital difference between eqs. (15) and (19). Equation (19) has been deduced from quantum mechanical considerations while equations like (15) are produced in a pure empirical context. But the interesting thing is the fact that eq. (19) somehow justifies eq. (15), when considering that QSAR-like parameters are nothing but rough approximations to QMSM or some appropriate transform.

The nature of the transformation (18) can be observed from many points of view. Two of them, among many possibilities, will be briefly described.

As a first example, let us suppose that the property or biological activity π appearing in eq. (19) has a macroscopic character; then eqs. (16) and (17), deduced in the quantum framework, are not so correct as in a microscopic environment. In this case the point-molecule u_I elements can be transformed in some statistical mechanics fashion into g_I elements, for instance, as

$$g_{JI} = \theta\exp[(u_{JI} - u_{II})/kT];\quad \forall J, I,\qquad\qquad(20)$$

where θ is some normalization constant.

The second example may serve to start the generalization of the molecular connectivity and related parameters. The main idea is based on the description and calculation algorithms of a new quantum related molecular topological descriptors parameter set [10]. It is possible to define the counterparts of many classical topological indices into the framework of the QMSM theory. For example, the elements of the topological matrix can be replaced by atomic ns shell orbital overlap integrals or more sophisticated measures like the ones described in section 2, tridimensional distances can be used instead of topological ones, effective charge parameters can enter into the definition of new indices, and so on. Essentially, MO QMSM as discussed in [1c] or the related molecular self-similarities may be used as good candidates to QSPR parameters, substituting other concepts of empirical origin. Then, these new quantum related topological indices have the possibility of

containing tridimensional information on the molecular structures and, also, chemical structure information. In this context, this kind of indices may be able to distinguish rotamers, conformers, etc., contrary to the classical ones which cannot.

As a final choice, the quantum molecular similarity indices (QMSI) will be described as a way to manipulate the original information on QMSM in the same way as done in eq. (18). QMSI have been described since the first paper on the subject [1a,1c], and discussed later on in [1m]. The most typical and well-known QMSI is the cosine-like one described by Carbó et al. [1a], which will produce a vector g_I such as

$$g_{JI} = u_{JI}(u_{JJ}u_{II})^{-1/2}; \quad \forall J. \tag{21}$$

QMSI may be used as QSPR parameters directly or suffer further transformations. A discussion of the QMSI relationships will be performed elsewhere.

8. Conclusions

Quantum molecular similarity measures establish a non-empirical theoretical basis where QSPR or QSAR can be justified as scientific procedures. Although QSPR has been a very useful tool since early times in chemistry, a proof of the appropriate theoretical foundations is lacking. The present work provides this need with a robust structure based on quantum chemical considerations.

The discrete representation of both an electronic density distribution and a convenient operator, connected with a quantum mechanical definition of the expectation value concept and, through this way, with the evaluation of molecular properties has been exposed.

References

[1] (a) R. Carbó, M. Arnau and L. Leyda, Int. J. Quant. Chem. 17 (1980) 1185.
(b) R. Carbó and C. Arnau, in: *Medicinal Chemistry Advances*, eds. de las F.G. Heras and S. Vega (Pergamon Press, Oxford, 1981).
(c) R. Carbó and Ll. Domingo, Int. J. Quant. Chem. 23 (1987) 517.
(d) R. Carbó and B. Calabuig: Comp. Phys. Commun. 55 (1989) 117.
(e) R. Carbó and B. Calabuig, in ref. [2k].
(f) R. Carbó and B. Calabuig, J. Mol. Struct. (Teochem) 254 (1992) 517.
(g) R. Carbó and B. Calabuig, J. Chem. Inf. Comp. Sci. 32 (1992) 600.
(h) R. Carbó and B. Calabuig, in: *Structure, Interactions and Reactivity*, ed. S. Fraga (Elsevier, Amsterdam, 1992).
(i) R. Carbó and B. Calabuig, Int. J. Quant. Chem. 42 (1992) 1681.
(j) R. Carbó and B. Calabuig, Int. J. Quant. Chem. 42 (1992) 1695.
(k) R. Carbó, B. Calabuig, E. Besalú and A. Martínez, Mol. Eng. 2 (1992) 43.
(l) R. Carbó and E. Besalú, *Proc. 1st Girona Seminar on Molecular Similarity* (Kluwer,

Dordrecht, 1993).

(m) R. Carbó, E. Besalú, B. Calabuig and L. Vera, Adv. Quant. Chem. 25 (1994) 253.

(n) R. Carbó, J. Mestres, M. Solà and E. Besalú, in: *Molecular Similarity*, ed. K. Sen, Topics in Current Chemistry 173 (Springer, Berlin, 1995) pp. 31–62.

(o) J. Mestres, M. Solà, M. Duran and R. Carbó, J. Comp. Chem. 15 (1994) 1113.

(p) M. Solà, J. Mestres, R. Carbó and M. Duran, J. Am. Chem. Soc. 116 (1994) 5909.

(q) M. Solà, J. Mestres, R. Carbó and M. Duran, J. Chem. Inf. Comp. Sci. 34 (1994) 1047.

[2] (a) D.L. Cooper and N.L. Allan, J. Chem. Soc. Faraday Trans. 83 (1987) 449.

(b) D.L. Cooper and N.L. Allan, J. Comp. Aided Mol. Des. 3 (1989) 253.

(c) D.L. Cooper and N.L. Allan, J. Am. Chem. Soc. 114 (1992) 4773.

(d) J. Cioslowski and E.D. Fleischmann, J. Am. Chem. Soc. 113 (1991) 64.

(e) J. Cioslowski and M. Challacombe, Int. J. Quant. Chem. 25 (1991) 81.

(f) J. Cioslowski and P.R. Surján, J. Mol. Struct. (Theochem) 255 (1992) 9.

(g) J.V. Ortiz and J. Cioslowski, Chem. Phys. Lett. 185 (1991) 270.

(h) R. Ponec and M. Strnad, Collect. Czech. Chem. Commun. 55 (1990) 2583.

(i) R. Ponec and M. Strnad, J. Phys. Org. Chem. 4 (1991) 701.

(j) R. Ponec and M. Strnad, Int. J. Quant. Chem. 42 (1992) 501.

(k) M.A. Johnson and G. Maggiora (eds.), *Concepts and Applications of Molecular Similarity* (Wiley, New York, 1990).

(l) E.E. Hodgkin and W.G. Richards, Int. J. Quant. Chem. 14 (1987) 105.

(m) A.C. Good, E.E. Hodgkin and W.G. Richards, J. Chem. Inf. Comp. Sci. 32 (1991) 188.

(n) M. Martín, F. Sanz, M. Campillo, L. Pardo, J. Pérez and J. Turmo, Int. J. Quant. Chem. 23 (1983) 1627.

(o) M. Martín, F. Sanz, M. Campillo, L. Pardo, J. Pérez, J. Turmo and J.M. Aulló, Int. J. Quant. Chem. 23 (1983) 1643.

(p) F. Sanz, M. Martín, J. Pérez, J. Turmo, A. Mitjana and V. Moreno, in: *Quantitative Approaches to Drug Design*, ed. J.C. Dearden (Elsevier, Amsterdam, 1983).

(q) F. Sanz, M. Martín, F. Lapeña and F. Manaut, Quant. Struct.-Act. Relat. 5 (1986) 54.

(r) F. Sanz, F. Manaut, J. José, J. Segura, M. Carbó and R. de la Torre, J. Mol. Struct. (Theochem) 170 (1988) 171.

(s) F. Manaut, F. Sanz, J. Jose and M. Milesi, J. Comp.-Aided Mol. Des. 5 (1991) 371.

(t) F. Sanz, F. Manaut, T. Dot and E. López de Briñas, J. Mol. Struct. (Theochem) 256 (1992) 287.

(u) F.J. Luque, F. Sanz, F. Illas, R. Pouplana and Y.G. Smeyers, Eur. J. Med. Chem. 23 (1988) 7.

[3] See, for an assorted set of early procedures:

(a) A. Crum-Brown and T. Fraser, Trans. Roy. Soc. Edinburgh 25 (1968-69) 151.

(b) E. Overton, Z. Physikol. Chem. 22 (1987) 189.

(c) H. Meyer, Arch. Exptl. Pathol. Pharmakol. 42 (1989) 109.

(d) J. Traube, Arch. Ges. Physiol. 105 (1904) 541.

(e) W. Moore, Science 49 (1919) 572.

(f) L.P. Hammet, Chem. Rev. 17 (1935) 125.

(g) J.C. McGowan, J. Appl. Chem. (London) 4 (1954) 41.

(h) C. Hansch and T. Fujita, J. Am. Chem. Soc. 86 (1964) 1616.

[4] See, for recent examples:

(a) J. Gálvez, R. García-Domenech, J.V. de Julián-Ortiz and R. Soler, J. Chem. Inf. Comp. Sci. 35 (1995) 272.

(b) M. Pastor and J. Alvarez-Bulla, Quant. Struct.-Act. Relat. 14 (1995) 24.

(c) M.D. Wessel and P.C. Jurs, J. Chem. Inf. Comp. Sci. 35 (1995) 68.

[5] R. Benigni, M. Cotta-Ramusino, F. Giorgi and G. Gallo, J. Med. Chem. 38 (1995) 629.

[6] (a) P.O. Löwdin, Phys. Rev. 97 (1955) 1474, 1490, 1509.
     (b) R. McWeeny, Proc. Roy. Soc. A232 (1955) 114.
     (c) R. McWeeny, Proc. Roy. Soc. A235 (1956) 496.
     (d) R. McWeeny, Proc. Roy. Soc. A253 (1959) 242.

[7] See, for example:
     (a) W.P. Purcell, G.E. Bass and John M. Clayton, *Strategy of Drug Design* (Wiley, New York, 1973).
     (b) L.B. Kier and L.H. Hall, *Molecular Connectivity in Chemistry and Drug Research* (Academic Press, New York, 1976).
     (c) W.G. Richards, *Quantum Pharmacology* (Butterworths, London, 1977).
     (d) Y.C. Martin, *Quantitative Drug Design*, Medicinal Research Series (Marcel Dekker, New York, 1978).
     (e) P. Krogsgaard-Larsen and H. Bundgaard (eds.), *A Textbook of Drug Design and Development* (Harwood, Chur, Switzerland, 1991).
     (f) A. Mosqueria (ed.), *Diseño de Medicamentos*, Real Academica de Farmacia, Madrid (1994).

[8] R. Carbó, M. Martín and V. Pons, Afinidad 34 (1977) 348.

[9] A. Bohm and M. Gadella, *Dirac Kets, Gamov Vectors and Gel'fand Triplets. The Rigged Hilbert Space Formulation of Quantum Mechanics*, Lecture Notes in Physics 348 (Springer, Berlin, 1989).

[10] E. Besalú and R. Carbó, Scientia Gerundensis 21 (1995).