Extending molecular similarity to energy surfaces: Boltzmann similarity measures and indices

Ramon Carbó-Dorca and Emili Besalú

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

Received 18 June 1996

Maxwell-Boltzmann statistics provides the adequate mathematical background allowing to define similarity measures involving molecular energy surfaces and electrostatic potential maps. Boltzmann similarity measures are described and various illustrative examples are used to show the practical viability of the theory. A new molecular similarity index is also presented. Finally, hybrid measures involving Boltzmann and density distributions are defined.

1. Introduction

During the past sixteen years our laboratory has initiated and developed the theoretical framework of Molecular Quantum Similarity (MQS) [1–4] as well as some auxiliary techniques have been described [5] and the whole applied to several chemical problems [2]. As a consequence, MQS Measures (MQSM) appear to be a fundamental tool [4a] in this field.

At the same time, various authors have presented the comparison of the Molecular Electrostatic Potential (MEP) maps between two molecules [6]. While the potential involved integral divergence was taken into account in some papers [1c,d], the rest of the literature did not mention this feature. Some studies circumvented the problem by using a linear combination of Gaussian functions approach [6d], which is almost the same as to transform the MEP into a density, while other authors tried to use statistical indexes [7] to do the job. Another answer was given recently by our laboratory [1o] in a MQSM general definition context, but this took into account the MEP electronic part only. In any case, the unsatisfactory computational pattern is evident when MQSM are tried, instead of density functions, in comparisons between potential energy surfaces. Still no complete procedure is operative and as far as our knowledge of the MQS state-of-the-art goes, up to now it has not yet been described.

In previous papers a clear statement of the correct definition of MQSM has been made [1k]. It was an insistent description of the need to use as MQSM basis sets *definite positive* functions, like those furnished by the density matrix elements. This was still clearly established when approximate, highly accurate density functions

were computed [3]. This was done due to the mathematical concept of measure, see ref. [8] for more details.

However, energy surfaces are not usually behaving like density matrices. Take a MEP function as a genuine example. A MEP has positive and negative regions, and due to the discontinuities present, it is not square summable without using appropriate weight functions [1c,d].

In the present paper we try to find a possible general solution to this problem: trying to somehow uniform the quantum and statistical mechanics definitions of similarity measures. First, we describe a simple recipe, then we deal with various naïve examples, while a final application of the previous theoretical results closes the discussion.

2. How similar is one energy surface to another?

So far MQSM deal with density functions and these, according to quantum mechanically well-stablished statements, are to be considered as statistical probability distributions. MQSM are nothing but a generalization of volumetric evaluations [8]. Similarity between energy surfaces must be defined coherently within a new framework bearing the same characteristics in MQSM problems.

Fortunately, there is an immediate solution to the problem: Boltzmann distributions [9] will transform non-definite energy surfaces into definite positive, square summable functions. Then one can deal with the comparison of statistical probability distributions and in this way it can be possible to define molecular similarity measures over *electronic energy surfaces* of any origin.

Therefore, let us set up the simple mathematical framework in the following way:

Let $E_A(\boldsymbol{r})$ be an energy surface attached to some state of an electronic system A. The vector $\boldsymbol{r}$ corresponds to the system's chosen degrees of freedom. A Boltzmann Partition Function (BPF) $p_A(\boldsymbol{r})$ [9] can be associated to the energy surface by using

$$p_A(\boldsymbol{r}) = \theta_A^{-1} \exp(-E_A(\boldsymbol{r})/KT)\,, \tag{1}$$

where K is the Boltzmann constant, T the temperature and θ_A is a normalization constant, defined as

$$\theta_A = \int_D \boldsymbol{W}(\boldsymbol{r}) \exp(-E_A(\boldsymbol{r})/KT)\, d\boldsymbol{r}\,, \tag{2}$$

$\boldsymbol{W}(\boldsymbol{r})$ being an optional weight operator attached to the domain of integration D. Such a weight function can be used to override the possible divergences of the previous integral, due to some particular forms of the energy surface: for instance, in some cases when $E_A(\boldsymbol{r}) \to 0$ if $|\boldsymbol{r}| \to \infty$.

Now, suppose that the partition function $p_B(\boldsymbol{r})$ is also known for a state of some

system B, then a similarity measure can be simply defined in accordance with the already described definitions of MQSM, which can be found in ref. [2d]:

$$B_{AB}(\Omega) = \int_{D_1}\int_{D_2} p_A(\boldsymbol{r}_1)\Omega(\boldsymbol{r}_1,\boldsymbol{r}_2)p_B(\boldsymbol{r}_2)\, d\boldsymbol{r}_1 d\boldsymbol{r}_2\,, \tag{3}$$

where $\Omega(\boldsymbol{r}_1,\boldsymbol{r}_2)$ is a weighting positive definite operator. Let us refer to the integrals $B_{AB}(\Omega)$ as Molecular Boltzmann Similarity Measures (MBSM).

A particular and very common case of the previous equation arises when using $\Omega(\boldsymbol{r}_1,\boldsymbol{r}_2) = \delta(\boldsymbol{r}_1 - \boldsymbol{r}_2)$, a Dirac delta function, as an operator in (3):

$$B_{AB} = \int_D p_A(\boldsymbol{r})p_B(\boldsymbol{r})\, d\boldsymbol{r}\,. \tag{4}$$

Next, in the same way as in the case of MQSM [4c], a pair, C-class and D-class, of similarity indices may be defined. For the C-class index,

$$C_{AB} = \frac{B_{AB}}{(B_{AA}B_{BB})^{1/2}}\,, \tag{5}$$

and for the D-class,

$$D_{AB} = (B_{AA} + B_{BB} - 2B_{AB})^{1/2}\,. \tag{6}$$

The index C_{AB} may be interpreted as a cosine-like function of the angle subtended by the BPFs $p_A(\boldsymbol{r})$ and $p_B(\boldsymbol{r})$, while D_{AB} corresponds to an Euclidean distance-like index between both functions seen as two points in an infinite-dimensional BPF space.

3. A normalized definition of similarity measures

The BPFs are normalized in the following sense:

$$\int_D W(\boldsymbol{r})p_A(\boldsymbol{r})\, d\boldsymbol{r} = \theta_A^{-1}\int_D W(\boldsymbol{r})\exp(-E_A(\boldsymbol{r})/KT)\, d\boldsymbol{r} = 1\,, \tag{7}$$

so the MBSM B_{AB} of eq. (4) may also be expressed by means of the integral

$$B_{AB} = \theta_A^{-1}\theta_B^{-1}\int_D \exp[-(E_A(\boldsymbol{r}) + E_B(\boldsymbol{r}))/KT]\, d\boldsymbol{r}\,, \tag{8}$$

in the same way as using normalized first order density functions in the realm of MQSM:

$$Z_{AB} = N_A^{-1}N_B^{-1}\int_D \rho_A(\boldsymbol{r})\rho_B(\boldsymbol{r})\, d\boldsymbol{r}\,, \tag{9}$$

where $\{\rho_A, \rho_B\}$ are density functions for systems A and B, respectively, and

$$N_I = \int_D \rho_I(\boldsymbol{r})\, d\boldsymbol{r}, \quad I = A, B, \tag{10}$$

is the number of electrons of system I found in the integration domain D.

It seems that the measure (9) will be a choice for MQSM coherent with the MBSM definition (4). Previously, a related MQSM has been used employing unit normalized density functions [3].

The normalization does not have any effect on the MBSM C-class index (5) nor on the one based on MQSM, though does have some influence on both the D-class index values.

4. Illustrative examples

In order to test the previous ideas let us propose various simple examples dealing with naïve energy functions.

(a) Vibrational potential functions

1. *Basic form.* Let us deal with a quadratic potential like αx^2, where α is a characteristic constant of system A involving the factor $(KT)^{-1}$, see below. The Boltzmann norm may be computed this time as the well-known integral [11]

$$\theta_A = \int_{-\infty}^{+\infty} \exp(-\alpha x^2)\, dx = \left(\frac{\pi}{\alpha}\right)^{1/2}, \tag{11}$$

where the weight operator W has been set to the unity. The MBSM is written by means of

$$\begin{aligned} B_{AB} &= (\theta_A \theta_B)^{-1} \int_{-\infty}^{+\infty} \exp(-(\alpha+\beta)x^2)\, dx = \pi^{-1/2}\left(\frac{\alpha\beta}{\alpha+\beta}\right)^{1/2} \\ &= (\pi KT)^{-1/2}\left(\frac{ab}{a+b}\right)^{1/2} \end{aligned} \tag{12}$$

if $\alpha = a(KT)^{-1}$ and $\beta = b(KT)^{-1}$, while the self-similarity measure for system A can be written as

$$B_{AA} = (\pi KT)^{-1/2}\left(\frac{a}{2}\right)^{1/2}, \tag{13}$$

with an equivalent expression for system B.

In this case, the C-class index will appear to be

$$C_{AB}^2 = \frac{(ab)^{1/2}}{\frac{1}{2}(a+b)}, \tag{14}$$

that is, the ratio between geometric and arithmetic means can be used to construct a good C-class index, comparable to more sophisticated cases [4c].

2. *Generalized measures.* One can define a weighted Boltzmann similarity measure as

$$B_{AB}(x^{2n}) = (\theta_A \theta_B)^{-1} \int_{-\infty}^{+\infty} x^{2n} \exp[-(\alpha+\beta)x^2]\, dx\,, \tag{15}$$

the partition functions being defined as in (11). The integral in (15) is well known to be [11]

$$\int_{-\infty}^{+\infty} x^{2n} \exp[-(\alpha+\beta)x^2]\, dx = \frac{n!!}{2^n(\alpha+\beta)^n}\left(\frac{\pi}{\alpha+\beta}\right)^{1/2}, \tag{16}$$

so, in this case,

$$B_{AB}(x^{2n}) = \frac{n!!}{2^n \pi^{1/2}}\left(\frac{\alpha\beta}{(\alpha+\beta)^{2n+1}}\right)^{1/2}, \tag{17}$$

and self-similarity will be defined as

$$B_{AA}(x^{2n}) = \frac{n!!}{2^{n+1}\pi^{1/2}}(2\alpha)^{(1-2n)/2}. \tag{18}$$

The corresponding C-class index is given by

$$C^2_{AB}(x^{2n}) = \left(\frac{(\alpha\beta)^{1/2}}{\frac{1}{2}(\alpha+\beta)}\right)^{2n+1}. \tag{19}$$

(b) *Torsional potential functions*

When the potential energy is defined as the torsional terms appearing in the usual formalism of molecular force fields [10],

$$E_A = U_A \cos(n\phi)\,, \tag{20}$$

where U_A and n are constants, the Boltzmann norm may be computed by using

$$\theta_A = \int_D W(\phi) \exp[-(U_A/KT)\cos(n\phi)]\, d\phi\,, \tag{21}$$

where the weighting operator depends on the angle ϕ, and the domain of integration D can be taken as the interval $[0, \pi/n]$.

For the case when $W(\phi) = 1$, it can be easily shown that the normalization constant is

$$\theta_A = \frac{\pi}{n} I_o\left(\frac{U_A}{KT}\right), \tag{22}$$

the function I_o being defined as

$$I_o(z) = \sum_{k=0}^{\infty} \left(\frac{z^k}{2^k k!} \right)^2 , \tag{23}$$

a zeroth order modified Bessel function of the first kind [11].

Another trivial case is obtained when $W(\phi) = \sin(n\phi)$; then the normalization constant becomes

$$\theta_A = \frac{2KT}{nU_A} \sinh\left(\frac{U_A}{KT} \right) . \tag{24}$$

(c) *Atomic electrostatic potential functions*

Suppose an atomic positive electrostatic potential defined as a Coulomb function:

$$E_A = Z_A |\boldsymbol{r}|^{-1} , \tag{25}$$

where Z_A is an atomic charge. Then a Boltzmann norm may be computed as

$$\theta_A = \int_D W(\boldsymbol{r}) \exp(-\gamma_A |\boldsymbol{r}|^{-1}) \, d\boldsymbol{r} , \tag{26}$$

where

$$\gamma_A = Z_A (KT)^{-1} , \tag{27}$$

and the weighting operator W must be defined in order to skip the divergence of the integrand when $|\boldsymbol{r}| \to \infty$.

In general, for energy potentials of the type $E(\boldsymbol{r}) = \kappa |\boldsymbol{r}|^{-n}$, κ being a constant, the weighting operators attached to the whole tri-dimensional space integration domain, $D = R^3$, can be defined in the form of $W(\boldsymbol{r}) = |\boldsymbol{r}|^{-m}$, with $m > 3$.

For the potential function analyzed here, a good choice is to take the weight operator defined as $W(\boldsymbol{r}) = |\boldsymbol{r}|^{-4}$, then the normalization constant becomes

$$\theta_A = 4\pi\gamma_A^{-1} . \tag{28}$$

It is easy to see that the attached BPF has the expression

$$p_A(\boldsymbol{r}) = \gamma_A^{-1} \exp(-\gamma_A |\boldsymbol{r}|^{-1}) . \tag{29}$$

By using a pair of BPFs like the one defined in eq. (29) a MBSM can be computed by evaluating the integral

$$B_{AB} = (\gamma_A \gamma_B)^{-1} 4\pi \int_0^{\infty} \Omega(|\boldsymbol{r}|) \exp[-|\boldsymbol{r}|^{-1} (\gamma_A + \gamma_B)/KT] \, d\boldsymbol{r} . \tag{30}$$

Choosing the operator $\Omega(|\boldsymbol{r}|) = |\boldsymbol{r}|^{-4}$ and reordering the terms, yields

$$B_{AB} = \frac{\gamma_A \gamma_B}{\gamma_A + \gamma_B} = (KT)^{-1} \frac{Z_A Z_B}{Z_A + Z_B}, \tag{31}$$

while self-similarities may easily be deduced as having the form

$$B_{II} = \frac{\gamma_I}{2} = (KT)^{-1} \frac{Z_I}{2}, \quad I = A, B. \tag{32}$$

So in this case, the C-class index is deduced to be

$$C_{AB} = \frac{(Z_A Z_B)^{1/2}}{\frac{1}{2}(Z_A + Z_B)}, \tag{33}$$

that is, the ratio of the geometric to the arithmetic mean of atomic charges is also found. From now on this index will be referred to as the *statistical index*. The statistical index form is equivalent to the one found in eq. (14) in the case of the Gaussian potential studied above. The origin of this feature can be found in the fact that $|\boldsymbol{r}|^{-1}$ may be expressed as a Gaussian transform of the type [11,12]

$$|\boldsymbol{r}|^{-1} = \pi^{-1/2} \int_{-\infty}^{+\infty} \exp(-|\boldsymbol{r}|^2 x^2)\, dx. \tag{34}$$

(d) *Lennard-Jones potential functions*

The ideas presented in the previous example can be applied to the study of Lennard-Jones (*q-p*) potential functions [9d,13], which can be written as

$$E(\boldsymbol{r}) = a|\boldsymbol{r}|^{-q} - b|\boldsymbol{r}|^{-p}, \tag{35}$$

where a and b are constants depending on the system under study.

A (*q-p*) Lennard-Jones potential function is more suitable to represent a realistic molecular energy function than the previous examples. In fact, this kind of potential has a form which resembles the shape of the electronic plus nuclear energy of a diatomic molecule bound state.

The Boltzmann norm in eq. (2) can be written in this case by means of the integral

$$\theta_A = 4\pi \int_0^{\infty} W(|\boldsymbol{r}|)|\boldsymbol{r}|^2 \exp(-\alpha_A |\boldsymbol{r}|^{-\lambda p} + \beta_A |\boldsymbol{r}|^{-p})\, dr, \tag{36}$$

where λ is a constant integer and α_A and β_A are the Lennard-Jones original constants divided by KT. Thus, the most usual values producing the well-known (12-6) Lennard-Jones potential are $\lambda = 2$ and $p = 6$. Choosing the weight function as $|\boldsymbol{r}|^{-(3+p)}$ and expanding as a power series the positive exponential part, after rearrangement the norm (36) can be given as

$$\theta_A = \frac{4\pi}{p\lambda} \sum_{l=0}^{\infty} \left(\frac{\beta_A^l \Gamma\left(\frac{l+1}{\lambda}\right)}{l! \alpha_A^{(l+1)/\lambda}} \right). \tag{37}$$

The Boltzmann measure (3) can be readily obtained in a similar way without further problem. The particular case with $\lambda = 2$ is worth mentioning because eq. (37) may be simplified due to the possibility of directly evaluating the corresponding norm (36) [11]:

$$\begin{aligned} \theta_A &= \frac{4\pi}{p} \int_0^{\infty} \exp(-\alpha_A x^2 + \beta_A x)\, dx \\ &= \frac{2\pi^{3/2}}{p\alpha_A^{1/2}} \exp\left(\beta_A^2/(4\alpha_A)\right) [1 + erf(\beta_A/(2\alpha_A^{1/2}))], \end{aligned} \tag{38}$$

$erf(x)$ being the error function.

5. A new similarity index: The statistical index

From the definition of MBSM it may be deduced that a suitable similarity index between two systems, A and B, of which some definite positive properties are known, $\{\alpha_A, \alpha_B\}$, say, can be defined as

$$C_{AB} = \varphi\left(\frac{(\alpha_A \alpha_B)^{1/2}}{\frac{1}{2}(\alpha_A + \alpha_B)} \right), \tag{39}$$

$\varphi(x)$ being some well-behaved function. For instance, in eq. (14) $\varphi(x) = x^{1/2}$; in eq. (19) one has $\varphi(x) = x^{(2n+1)/2}$, while in eq. (33) $\varphi(x) = x$.

In this sense, one can say that multiple system comparisons may be performed by means of the appropriate ratios between geometric means:

$$\gamma = \left(\prod_{i=1}^{n} \alpha_i \right)^{1/n}, \tag{40}$$

and arithmetic ones:

$$\mu = \frac{1}{n} \sum_{i=1}^{n} \alpha_i, \tag{41}$$

as

$$C_{AB} = \varphi\left(\frac{\gamma}{\mu} \right). \tag{42}$$

The ratio in eq. (42) will tend to unity when all the applied systems are equal and will become zero when one of the properties, α_k, say, goes to infinity.

This is so because if $\alpha_k \gg \alpha_i, \forall i \neq k$, then

$$\lim_{\alpha_k \to \infty} \frac{\gamma}{\mu} \simeq \left[\prod_{i \neq k} (\alpha_i)^{1/n}\right] \frac{\alpha_k^{1/n}}{\frac{1}{n}\alpha_k} = \left[\prod_{i \neq k} (\alpha_i)^{1/n}\right] n \alpha_k^{(1-n)/n} \propto \alpha_k^{-1} = 0 . \tag{43}$$

The form of the C-class statistical index, as defined in eq. (33), resembles the one proposed by Hodgkin and Richards [6c] in the MQS framework. Then, as it was shown recently that Carbó and Hodgkin-Richards indices are related [4c], it must be expected that it is also related to the Carbó index.

As an application example, Tables 1 to 3, reproduced for three different molecular families: 13 feromones [14] (Table 1), 9 fluoro and chloromethanes [4d] (Table 2) and the first 16 linear alkanes (Table 3), contain the mean quadratic differences (lower triangle) and correlation indices (upper triangle) between various matrices of C-class indices: Carbó, Hodgkin-Richards, Tanimoto, Petke, a new index defined recently [4c] that takes into account the discrete representation of density functions, and the statistical index. This last index is obtained using as parameters the molecular self-similarity measures, that is: $Z_A = Z_{AA}$ and $Z_B = Z_{BB}$. Also, the function in the statistical index definition (39) has been taken to be $\varphi(x) = x$.

From the inspection of the three tables it can be commented that, in agreement with ref. [4c], Carbó and Hodgkin-Richards indices are almost the same as well as the Petke and Tanimoto pair of indices, which present an even more close relationship. Curiously enough, it can be seen how the statistical index values differ substantially from the Hodgkin-Richards, Tanimoto and Petke indices, while they are more similar to the Carbó index and are even closer to the index defined in ref. [4c], following the fact that the discrete and the statistical indices present the most accused differences with the rest. These features are repeated in all the studied cases.

It must be finally said that the statistical index discussed here can be used in a theoretical field of molecular parameters, as well as in an empirical set of experimental ones, becoming in this way a general tool for molecular similarity purposes. As the

Table 1
Mean quadratic error (lower triangle) and correlation coefficient (upper triangle) between the matrices containing C-class indices computed for a family of 13 feromones [14]. CAR, HR, TAN, PET, DC and S stand for the Carbó, Hodgkin-Richards, Tanimoto, Petke, the index defined in ref. [4c] and the statistical index, respectively.

Index	CAR	HR	TAN	PET	DC	S
CAR		0.964377	0.990046	0.960522	0.608673	0.603641
HR	0.113680		0.980275	0.996425	0.394171	0.791545
TAN	0.186414	0.124184		0.979306	0.507425	0.669886
PET	0.147264	0.501806E-01	0.907430E-01		0.387813	0.788576
DC	0.365056	0.463141	0.549909	0.502976		−0.215457
S	0.356485	0.375519	0.495207	0.414764	0.336264	

Table 2
Mean quadratic error (lower triangle) and correlation coefficient (upper triangle) between the matrices containing C-class indices computed for a family of 9 fluoro and chloromethanes. CAR, HR, TAN, PET, DC and S stand for the Carbó, Hodgkin-Richards, Tanimoto, Petke, the index defined in ref. [4c] and the statistical index, respectively.

Index	CAR	HR	TAN	PET	DC	S
CAR		0.982560	0.962126	0.931322	0.918445	0.418522
HR	0.560301E-01		0.984943	0.979310	0.872719	0.571239
TAN	0.180902	0.138866		0.983677	0.798608	0.553073
PET	0.158297	0.106957	0.606572E-01		0.779086	0.670202
DC	0.267756	0.308416	0.443481	0.409545		0.272597
S	0.397070	0.425594	0.553890	0.513894	0.175527	

parameters, needed to construct the index, come individually from every molecule in the compared set, a mixture of theoretical and empirical values, taking into account possible sign alternations, is also feasible within the statistical index definition.

6. Hybrid similarity measures and QSAR

In a recent paper, a theoretical background connection between MQSM and QSAR was stablished [4b]. A complete equivalent formalism can be applied here on the set of MBSM. But here is a new possibility which may connect both kinds of similarity measures and, thus, enlarge the possibilities of explaining the success of structure-activity or structure-properties relationships.

In the same way as in the MQSM and MBSM computation, suppose a set of molecules $M = \{m_I\}$, the attached set of densities normalized to one particle $P = \{\rho_I(\boldsymbol{r})\}$ and a set of energy surfaces $E = \{E_I(\boldsymbol{R})\}$, which can be used to construct a set of Boltzmann partition functions $B = \{p_I(\boldsymbol{R})\}$ by using

Table 3
Mean quadratic error (lower triangle) and correlation coefficient (upper triangle) between the matrices containing C-class indices computed for a family of 16 linear alkanes. CAR, HR, TAN, PET, DC and S stand for the Carbó, Hodgkin-Richards, Tanimoto, Petke, the index defined in ref. [4c] and the statistical index, respectively.

Index	CAR	HR	TAN	PET	DC	S
CAR		0.999563	0.988464	0.988463	0.977684	0.933585
HR	0.671489E-01		0.985835	0.985835	0.982039	0.937491
TAN	0.184479	0.128155		1.00000	0.936706	0.868793
PET	0.184485	0.128162	0.449830E-04		0.936704	0.868793
DC	0.265055	0.330410	0.446532	0.446538		0.980182
S	0.214783	0.277085	0.399194	0.399200	0.716913E-01	

$$p_I(\boldsymbol{R}) = \theta_I^{-1} \exp(-E_I(\boldsymbol{R})/KT) \tag{44}$$

with

$$\theta_I = \int_D \boldsymbol{W}(\boldsymbol{R}) p_I(\boldsymbol{R})\, d\boldsymbol{R}\,. \tag{45}$$

A correspondence between the sets P, M and B,

$$\rho_I \leftrightarrow m_I \leftrightarrow p_I\,, \tag{46}$$

is thus defined. Then, two kinds of Hybrid Similarity Measures (HSM) may be defined by computing the integrals

$$f_{IJ}(\Omega) = \int\int p_I(\boldsymbol{R}) \Omega(\boldsymbol{R}, \boldsymbol{r}) \rho_J(\boldsymbol{r})\, d\boldsymbol{R} d\boldsymbol{r} \tag{47}$$

and

$$g_{IJ}(\Theta) = \int\int \rho_I(\boldsymbol{r}) \Theta(\boldsymbol{r}, \boldsymbol{R}) p_J(\boldsymbol{R})\, d\boldsymbol{r} d\boldsymbol{R}\,, \tag{48}$$

where $\Omega(\boldsymbol{R}, \boldsymbol{r})$ and $\Theta(\boldsymbol{r}, \boldsymbol{R})$ are appropriate operators as defined in eq. (9). The corresponding matrices will not be symmetrical, contrary to the pure cases.

Let us now associate to every molecule a vector by means of MQSM and MBSM projections of the sets P and B into an n-dimensional vector space:

$$\rho_I \rightarrow \boldsymbol{z}_I = \{z_{JI}\} \tag{49}$$

and

$$p_I \rightarrow \boldsymbol{b}_I = \{b_{JI}\}\,. \tag{50}$$

In this manner, a new relationship may be envisaged,

$$\boldsymbol{z}_I \leftrightarrow m_I \leftrightarrow \boldsymbol{b}_I\,, \tag{51}$$

as an n-dimensional representation of the set M as a discrete alternative to the continuous representation (46). However, the matrix elements formed by the integrals (47) and (48) correspond to a new discrete molecular representation involving the Cartesian products $B \otimes P$ and $P \otimes B$, respectively, instead of the usual $P \otimes P$ and $B \otimes B$ attached to the pure MQSM and MBSM. One can thus speak of another set of molecular representations made by the rows or columns of matrices $\boldsymbol{F} = \{f_{IJ}\}$ and $\boldsymbol{G} = \{g_{IJ}\}$. For example, by using a column partition $\boldsymbol{F} = (\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_n)$ and $\boldsymbol{G} = (\boldsymbol{g}_1, \boldsymbol{g}_2, \ldots, \boldsymbol{g}_n)$, one obtains

$$\rho_I \rightarrow \boldsymbol{f}_I \quad \wedge \quad p_I \rightarrow \boldsymbol{g}_I\,, \tag{52}$$

and a reverse relationship when considering rows.

When the operators of HSM are the same, $\Omega = \Theta$, then the hybrid measures are related by means of $f_{IJ} = g_{JI}$, $\forall I, J$, so $\boldsymbol{F} = \boldsymbol{G}^{\mathrm{T}}$, and provided that the involved

operators in similarity measures are definite positive and thus Hermitian, a symmetric HSM can be defined using the symmetrizer $\boldsymbol{J} = (\boldsymbol{F} + \boldsymbol{G})/2$.

HSM may generalize the theoretical background discussed some time ago in [4b]. An integral like (47) may be used in a possible extension of the QSAR equation:

$$\pi_I = \langle \Omega \rangle_I = \langle \Omega \mid \rho_I \rangle \sim \boldsymbol{w}^T \boldsymbol{z}_I \,, \tag{53}$$

where π_I is some property for molecule m_I and $\boldsymbol{w}$ is an unknown operator representation of the same dimensionality as density ρ_I represented by $\boldsymbol{z}_I$. Now, eq. (53) can be rewritten with the help of HSM in terms of the previous unknown operator average over a Boltzmann partition function, $p(\boldsymbol{R})$, that is,

$$\pi_I = \int \int p(\boldsymbol{R}) \Omega(\boldsymbol{R}, \boldsymbol{r}) \rho_I(\boldsymbol{r}) d\boldsymbol{R}\, d\boldsymbol{r} = \int w(\boldsymbol{r}) \rho_I(\boldsymbol{r})\, d\boldsymbol{r} \sim \boldsymbol{w}^T \boldsymbol{z}_I \,, \tag{54}$$

where the new operator $\boldsymbol{w}$ is defined as

$$w(\boldsymbol{r}) = \int p(\boldsymbol{R}) \Omega(\boldsymbol{R}, \boldsymbol{r})\, d\boldsymbol{R} \,. \tag{55}$$

This produces a very general possible interpretation for the linear coefficients in QSAR and QSPR least squares equations.

As an example of a HSM, here we present the application to the ground state of a hydrogenoid atom using a nuclear electrostatic potential as an energy surface. The normalized density function is defined as

$$\rho(\boldsymbol{r}) = \frac{Z^3}{\pi} \exp(-2Z\boldsymbol{r}) \,, \tag{56}$$

Z being the atomic charge. The Boltzmann partition function is defined as

$$p(\boldsymbol{r}) = \frac{4\pi KT}{Z} \exp(-\gamma |\boldsymbol{r}|^{-1}) \,, \tag{57}$$

where $\gamma = Z(KT)^{-1}$. The evaluation of integral (47) gives different results depending on the form of the operator Ω. The three chosen immediate results for the hybrid self-similarity measure follow:

(a) for $\Omega = |\boldsymbol{r}|^{-3/2}$

$$f(|\boldsymbol{r}|^{-3/2}) = \eta \frac{1 + 2Z\delta}{4Z^{3/2}} \left(\frac{\pi}{2}\right)^{1/2} \exp(-2Z\delta) \,; \tag{58}$$

(b) for $\Omega = |\boldsymbol{r}|^{-5/2}$

$$f(|\boldsymbol{r}|^{-5/2}) = \eta \left(\frac{\pi}{2Z}\right)^{1/2} \exp(-2Z\delta) \,; \tag{59}$$

(c) for $\Omega = |r|^{-3}$

$$f(|r|^{-3}) = \eta 2K_0(2Z\delta)\,, \tag{60}$$

where $\eta = 16\pi KTZ^2$, $\delta = (2/(KT))^{1/2}$ and K_0 is the zeroth order modified Bessel function of the second kind [11].

7. Conclusions

Boltzmann similarity measures related to molecular energy surfaces have been described. Although in the practical cases, BSM based on molecular energy surfaces will necessarily need numerical integration techniques, it has been shown that for simple energy functions, related to electronic, vibrational and torsional phenomena, a new set of similarity measures and indices can be easily computed. A new similarity index, the statistical index: the ratio between geometric and arithmetic means of some set of molecular properties or parameters, has also been described and compared with well-known indices. QSAR and molecular similarity theory can be connected through MQSM and MBSM hybrid measures. The naïve mathematical framework developed in this paper fills a gap in the general structure of molecular similarity measures and starts the way to a complete theory on this subject.

Acknowledgements

This work was partially supported by a CICYT grant #SAF96-0158-C02-01. Results furnished by Mr. X. Fradera and Mr. Ll. Amat of the IQC-UdG, which have been used to construct the index comparisons, are greatly acknowledged.

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. F.G. de las Heras and S. Vega (Pergamon Press, Oxford, 1981).
(c) R. Carbó, E. Suñé, F. Lapeña and J. Pérez, J. Biological Phys. 14 (1986) 21.
(d) R. Carbó, F. Lapeña and E. Suñé, Afinidad 43 (1986) 483.
(e) R. Carbó and Ll. Domingo, Int. J. Quant. Chem. 23 (1987) 517.
(f) R. Carbó and B. Calabuig, Comp. Phys. Commun. 55 (1989) 117.
(g) R. Carbó and B. Calabuig, in: *Concepts and Applications of Molecular Similarity*, eds. M.A. Johnson and G. Maggiora (Wiley, New York, 1990) p. 147.
(h) R. Carbó and B. Calabuig, J. Mol. Struct. (Theochem) 254 (1992) 517.
(i) R. Carbó and B. Calabuig, J. Chem. Inf. Comput. Sci. 32 (1992) 600.
(j) R. Carbó and B. Calabuig, in: *Structure, Interactions and Reactivity*, ed. S. Fraga (Elsevier, Amsterdam, 1992).

(k) R. Carbó and B. Calabuig, Int. J. Quant. Chem. 42 (1992) 1681.
(l) R. Carbó and B. Calabuig, Int. J. Quant. Chem. 42 (1992) 1695.
(m) R. Carbó, B. Calabuig, E. Besalú and A. Martínez, Molec. Eng. 2 (1992) 43.
(n) R. Carbó and E. Besalú, *Proc. 1st Girona Seminar on Molecular Similarity* (Kluwer, Dordrecht, 1993).
(o) R. Carbó, E. Besalú, B. Calabuig and L. Vera, Adv. Quant. Chem. 25 (1994) 253.

[2] (a) J. Mestres, M. Solà, M. Duran and R. Carbó, J. Comp. Chem. 15 (1994) 1113.
(b) M. Solà, J. Mestres, R. Carbó and M. Duran, J. Am. Chem. Soc. 116 (1994) 5909.
(c) M. Solà, J. Mestres, R. Carbó and M. Duran, J. Chem. Inf. Comp. Sci. 34 (1994) 1047.
(d) R. Carbó, J. Mestres, M. Solà and E. Besalú, in: *Topics in Current Chemistry*, ed. K. Sen, Vol. 173, *Molecular Similarity* (Springer, Berlin, 1995) pp. 31–62.

[3] (a) P. Constants and R. Caró, J. Chem. Inf. Comput. Sci. 35 (1995) 1046.
(b) P. Constants, Ll. Amat, X. Fradera and R. Carbó, Quantum molecular similarity measures (QMSM) and atomic shell approximation (ASA), *Proc. 2nd Girona Seminar on Molecular Similarity*, eds. R. Carbó and P.G. Mezey, *Advances in Molecular Similarity*, Vol. 1 (JAI Press, Greenwich, Conn., in press).

[4] (a) R. Carbó, E. Besalú, Ll. Amat and X. Fradera, Quantum molecular similarity measures: Concepts, definitions and applications, *Proc. 2nd Girona Seminar on Molecular Similarity*, eds. R. Carbo and P.G. Mezey, *Advances in Molecular Similarity*, Vol. 1 (JAI Press, Greenwich, Conn., in press).
(b) R. Carbó, E. Besalú, L. Amat and X. Fradera, J. Math. Chem. 18 (1995) 237.
(c) R. Carbó, E. Besalú, Ll. Amat and X. Fradera, J. Math. Chem. 19 (1996) 47.
(d) R. Carbó and E. Besalú, Afinidad 53 (1996) 77.

[5] (a) R. Carbó and E. Besalú, Comput. & Chem. 18 (1994) 117.
(b) R. Carbó and E. Besalú, J. Math. Chem. 18 (1995) 37.
(c) R. Carbó and E. Besalú, in: *Strategies and Applications in Quantum Chemistry: From Astrophysics to Molecular Engineering*, eds. M. Defranceschi and Y. Ellinger (Kluwer, Amsterdam, 1996) pp. 229–248.

[6] (a) A.C. Good, S.S. So and W.G. Richards, J. Med. Chem. 36 (1993) 433.
(b) A.C. Good, S.J. Peterson and W.G. Richards, J. Med. Chem. 36 (1993) 2929.
(c) E.E. Hodgkin and W.G. Richards, Int. J. Quant. Chem. 14 (1987) 105.
(d) A.C. Good, E.E. Hodgkin and W.G. Richards, J. Chem. Inf. Comput. Sci. 32 (1992) 188.

[7] (a) M. Martín, F. Sanz, M. Campillo, L. Pardo, J. Pérez and J. Turmo, Int. J. Quant. Chem. 23 (1983) 1627.
(b) M. Martín, F. Sanz, M. Campillo, L. Pardo, J. Pérez, J. Turmo and J.M. Aulló, Int. J. Quant. Chem. 23 (1983) 1643.
(c) 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).
(d) F. Sanz, M. Martín, F. Lapeña and F. Manaut, Quant. Struct. Act. Relat. 5 (1986) 54.
(e) F. Sanz, F. Manaut, J. José, J. Segura, M. Carbó and R. de la Torre, J. Mol. Struct. (Theochem) 170 (1988) 171.
(f) F.J. Luque, F. Sanz, F. Illas, R. Pouplana and Y.G. Smeyers, Eur. J. Med. Chem. 23 (1988) 7.
(g) F. Manaut, F. Sanz, J. Jose and M. Milesi, J. Comput. -Aided. Mol. Des. 5 (1991) 371.
(h) F. Sanz, F. Manaut, T. Dot and E. López de Briñas, J. Mol. Struct. (Theochem) 256 (1992) 287.

[8] (a) *Encyclopaedia of Mathematics* (Kluwer, Dordrecht, 1990).
(b) A.N. Kolmogorov and S.V. Fomin, *Elementos de la Teoria de Funciones y del Analisis Funcional* (Mir, Moscow, 1975).
(c) A.E. Taylor, *General Theory of Functions and Integration* (Dover, New York, 1985).

(d) E.R. Philips, *An Introduction to Analisys and Integration Theory* (Dover, New York, 1984).

[9] (a) H. Eyring, D. Henderson, B.J. Stover and E.M. Eyring, *Statistical Mechanics and Dynamics* (Wiley, New York, 1964).
(b) G.H. Wannier, *Statistical Physics* (Dover, New York, 1966).
(c) A. Pacault, *Éléments de Thermodynamique Statistique* (Masson & Cie., Paris, 1963).
(d) J.O. Hirschfelder, Ch.F. Curtiss and R.B. Bird, *Molecular Theory of gases and Liquids* (Wiley, New York, 1964).

[10] D.M. Hirst, *A Computational Approach to Chemistry* (Blackwell, Oxford, 1990) see chap. 3 and references therein.

[11] (a) M. Abramowitz and I.A. Stegun (eds.), *Handbook of Mathematical Functions* (Dover, New York, 1965).
(b) W. Gröbner and N. Hofreiter, *Integraltafel* (Springer, Vienna, 1966).

[12] S. Obara and A. Saika, J. Chem. Phys. 84 (1986) 3963.

[13] J.E. Lennard-Jones, Proc. Roy. Soc. A106 (1924) 463.

[14] J.E. Amoore, *Molecular Basis of Odor* (Thomas, Springfield, 1970).