Molecular Quantum Similarity Measures Tuned 3D QSAR: An Antitumoral Family Validation Study†

Lluís Amat, David Robert, Emili Besalú, and Ramon Carbó-Dorca*

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

Received January 30, 1998

In this work, a new methodology to construct a tuned QSAR model is presented, which is based on a convex set formalism. The present procedure continues previous 3D QSAR studies, performed using molecular quantum similarity measures (MQSM). With this new computational tool, the efficiency of MQSM applied to QSAR analysis is significantly improved. A reliable QSAR model is obtained using convex linear combinations of different kinds of MQSM, corresponding to different quantum-mechanical operators related to the quantum similarity integral. The active compounds studied here, as a case study, are a set of antitumor agents, the camptothecin molecule and analogues, and the property evaluated is the topoisomerase-I inhibition activity. Before performing a tuned QSAR analysis with this particular molecular set, a simple QSAR study for all the different possible types of MQSM is carried out. In addition, another application of MQSM is presented, to determine which method can be used to optimize molecular structures in order to reproduce experimental molecular geometries as well as possible.

INTRODUCTION

Molecular similarity procedures play an important role in the success of structure-based and computer-aided drug design.[1−6] In recent years, among these techniques, molecular quantum similarity measures (MQSM) have become a useful tool for the determination of 3D QSAR[7−9] and could have a role of their own in the search for new biologically active compounds. Recent advances in computational algorithms and methodological procedures related to MQSM permit fast and appropriate performance of QSAR studies. The most relevant points associated to this possibility are as follows: (1) Construction of accurate fitted electronic density functions, with the so-called *atomic shell approximation*,[10−12] allowing inexpensive studies of large molecules to be carried out. (2) A new maximization algorithm to obtain optimal molecular superpositions.[13,14] (3) The quantum-mechanical formalism to connect MQSM with QSPR-QSAR analysis.[7] (4) The recent application of convex set concepts to obtain a general mathematical pattern, enveloping the MQSM framework,[15,16] giving as a consequence the possibility of combining different MQSM to produce new problem adapted measures.

The present study is related to this previous work, and as a result presents an application-example of tuned QSAR analysis using MQSM. The molecular set studied here is composed of the camptothecin (CPT) molecule and 11 analogues. These compounds are cytotoxic drugs which present a potent inhibitory function to nucleic acid synthesis in mammalian cells and induce the strand DNA breaks in reactions containing purified mammalian DNA topoisomerase-I (T-I).[17] In this paper, a QSAR analysis will be performed in order to predict the T-I inhibition IC_{50} values for the CPT and analogues. Structures and activities for these drugs have been reported by M. E. Wall and M. C. Wani.[18]

The procedure proposed here can be resumed as follows. In a first stage, a wide quantum similarity study for the set of 12 antitumor agents is carried out, computing different kinds of MQSM. After an individual study of each MQSM, and using a convex set formalism, a tuned 3D QSAR model is constructed, using positive definite linear combinations of the different MQSM, which can be used as optimal descriptors.

METHODS: THEORETICAL ASPECTS

Before presenting QSAR results, a brief description of the main theoretical aspects used in this work will be given. More detailed information about these procedures can be found in various recent published books,[19−22] where an extensive review of the theoretical aspects attached to MQSM is described.

Quantum Similarity Measures. In order to quantify the degree of similarity between two molecules, it is necessary to choose a molecular descriptor to calculate similarity measures. The MQSM considered here are based on quantum-mechanical postulates and use the first order electronic density function as molecular descriptor, which provides a theoretically coherent 3D molecular representation. Then, a MQSM between two molecules *A* and *B* is expressed by means of the following integral

$$Z_{AB}(\Omega_\alpha) = \int\int \rho_A(\mathbf{r}_1)\Omega_\alpha(\mathbf{r}_1,\mathbf{r}_2)\rho_B(\mathbf{r}_2)\, d\mathbf{r}_1\, d\mathbf{r}_2 \quad (1)$$

where $\{\rho_A(\mathbf{r}_1), \rho_B(\mathbf{r}_2)\}$ are the density functions of each molecule, and $\Omega_\alpha(\mathbf{r}_1,\mathbf{r}_2)$ is a positive definite operator. The subindex α denotes different kinds of possible MQSM definitions. When both molecules are the same, the related

* To whom correspondence should be addressed.

† Keywords: atomic shell approximation, camptothecin analogues, Carbó index, classical scaling analysis, convex sets, DNA topoisomerase-I, molecular quantum similarity measures, promolecular densities, similarity matrices, tuned 3D QSAR.

S0095-2338(98)00010-9 CCC: $15.00
Published on Web 06/03/1998

MQSM is denoted as a self-similarity measure, $\{Z_{AA},Z_{BB}\}$. The most common MQSM is defined substituting $\Omega_\alpha(\mathbf{r}_1,\mathbf{r}_2)$ by a Dirac delta function, $\delta(\mathbf{r}_1\text{-}\mathbf{r}_2)$, and then eq 1 is simplified to the so-called overlap-like MQSM:

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

Other operators used in this work are the Coulomb-like MQSM defined as $\Omega_\alpha(\mathbf{r}_1,\mathbf{r}_2) = |\mathbf{r}_1 - \mathbf{r}_2|^{-1}$, a gravitational-like MQSM, where $\Omega_\alpha(\mathbf{r}_1,\mathbf{r}_2) = |\mathbf{r}_1 - \mathbf{r}_2|^{-2}$ is used, and finally, a Triple MQSM (TMQSM), where the operator Ω_α is substituted by another molecular density, $\rho_C(\mathbf{r})$, transforming eq 1 into

$$Z_{AB;C} = \int \rho_A(\mathbf{r})\rho_C(\mathbf{r})\rho_B(\mathbf{r})\, d\mathbf{r} \tag{3}$$

Transformations of the MQSM produce new discrete descriptors, which are known as quantum similarity indices.[23] One of the most common ones found in the literature is the so-called Carbó index,[24] defined as

$$C_{AB} = Z_{AB}(Z_{AA}Z_{BB})^{-1/2} \tag{4}$$

which can be generalized in the TMQSM case as

$$C_{AB;C} = Z_{AB;C}(Z_{AA;C}Z_{BB;C})^{-1/2} \tag{5}$$

Fitted Density Functions: Atomic Shell Approximation. In order to avoid cumbersome *ab initio* calculations of density functions, some fitting algorithms have been developed to construct accurate density functions.[10−12,25,26] The *atomic shell approximation* (ASA)[10−12] is a well-established algorithm providing fitted density functions using a spherical 1*S-GTO* basis set. This approximation constructs the fitted density function using positive definite coefficients only and consequently preserves the statistical meaning of the density function.[27] Recent developments in ASA proposed a new algorithm to obtain a positive definite coefficient representation, based on the *elementary Jacobi rotations technique*.[12] In the same paper, a fast *promolecular* ASA was developed, corresponding to the expression

$$\rho_A^{\mathrm{ASA}}(\mathbf{r}) = \sum_{a\in A} P_a \rho_a^{\mathrm{ASA}}(\mathbf{r}) \tag{6}$$

where the sum in eq 6 runs over all the atoms of molecule *A*. Here, P_a has been chosen as the atomic number of each atom *a*, and atomic density function ρ_a^{ASA} is constructed using squared-normalized 1*S-GTO* functions centered on the *a*th atom

$$\rho_a^{\mathrm{ASA}}(\mathbf{r}) = \sum_{i\in a} \mathrm{w}_i |\mathrm{S}_i(\mathbf{r} - \mathbf{r}_a;\zeta_i)|^2 \tag{7}$$

while keeping the convex coefficient constraints:

$$\{\mathrm{w}_i > 0;\forall i\} \wedge \{\sum_{i\in a} \mathrm{w}_i = 1\} \tag{8}$$

In this way the total number of electrons of molecule *A* is obtained when density function ρ_A^{ASA} is integrated over all the space.

Molecular Superposition and Maximization Process. Optimal molecular superpositions are obtained using a compact and robust algorithm, based on an exact solution, if density functions are deformed to Dirac delta functions.[13,14] This algorithm constitutes a very suitable method when ASA density functions are considered. Two processes compose the superposition search methodology. In the first stage, a search over all the triplets of atoms of both molecules is performed until an optimal superposition is obtained. This process requires $n_A^3 n_B^3$ evaluations of the similarity function, where *n* is the number of atoms for each studied molecule. Different algorithm levels have been developed in order to simplify and accelerate the search procedure, reducing the number of evaluations of Z_{AB} to $n_A n_B$ and obtaining the same results as in the complete search.[13,14] After this search, and starting from the best molecular alignment which has been found, a new restricted search is implemented to refine the solution and obtain the maximal value of the similarity measure. In this process, a Newton or a Simplex technique is employed, depending on whether the first and second derivatives of a given MQSM are known.

TMQSM presents a special maximization, which in this study has been solved using bimolecular superpositions. The measure $Z_{AB;C}$ is optimized in the following way: while maintaining the molecule-operator *C* fixed in the space, overlap MQSM: Z_{CA}, of *A* with respect to *C*, and overlap MQSM: Z_{CB}, of *B* with respect to *C*, are optimized separately. After that, a simplex method is used with 12 variables (three rotations and three translations for both free molecules *A* and *B*) maximizing the $Z_{AB;C}$ measure. When studied molecules have a common structure, a satisfactory superposition of three molecules is obtained using the procedure described succinctly above.

Connection between QSAR and MQSM. After computing the MQSM for all the molecular pairs involved in the molecular set studied, **M**, a symmetric matrix is obtained, which is known as the similarity matrix, **Z**. This square ($m \times m$) matrix **Z**, where *m* is the number of molecules contained in the set **M**, can be considered as a row hypermatrix whose elements are *m*-dimensional column vectors, collecting all the matrix elements associated to a given molecule:[28]

$$\mathrm{Z}(\Omega_\alpha) = \{\mathbf{z}_1,\mathbf{z}_2,...\mathbf{z}_m\} \wedge \mathbf{z}_I = \{\mathrm{Z}_{JI};\, \forall\, J \in \mathbf{M}\} \tag{9}$$

In this expression, Ω_α denotes the different kinds of MQSM, corresponding to the different weight operators which can be used in eq 1. Vectors $\mathbf{z}_I$ can be interpreted as the matrix representation of each molecule in the vector space spanned by the density functions weighted with the Ω_α operator. Within this definition, each vector $\mathbf{z}_I$ has been called a point-molecule belonging to a molecular point cloud, represented by the collection of columns of **Z**.

In a recent work,[7] a relationship between molecular properties, π_I, and the discrete representation of molecular descriptors defined by the vectors $\mathbf{z}_I$ was deduced, producing a linear equation

$$\pi_I = \mathbf{b}^{\mathrm{T}}\mathbf{z}_I \tag{10}$$

where **b** is an *m*-dimensional vector, whose elements must be determined. The most usual way of computing **b** elements is a least-squares technique.

The main statistical parameters used to evaluate model accuracy are the square regression coefficient (r^2) and the square predictive regression coefficient (Q^2). This last statistical coefficient is computed using the expression

$$Q^2 = 1 - \frac{\text{PRESS}}{\sum_I (\pi_{I,\text{obs}} - \bar{\pi})^2} \tag{11}$$

and defining the predictive residual sum of squares (PRESS) as

$$\text{PRESS} = \sum_I \left(\frac{\pi_{I,\text{obs}} - \pi_{I,\text{calc}}}{1 - h_{II}} \right)^2 \tag{12}$$

where $\pi_{I,\text{obs}}$ and $\pi_{I,\text{calc}}$ are the observed and calculated property values for the compound I, $\bar{\pi}$ is the averaged property values, and h_{II} are the diagonal elements of the so-called Hat matrix (see ref 29).

Convex Set Formalism Applied to MQSM. The term convex set can be interpreted as a collection of linear combinations of vectors in a vector semispace,[15,16] where the coefficients fulfill the constraints defined in eq 8. The generated vectors using convex constraints can be considered normalized and contained within the unit sphere. In the present study, convex sets have been associated to the manipulation of some molecular information included in the similarity matrices, $\mathbf{Z}(\Omega_\alpha)$, by means of a positive definite linear combination

$$\mathbf{Z} = \sum_{\alpha=1}^{\upsilon} c_\alpha \mathbf{Z}(\Omega_\alpha) \tag{13}$$

where the set of $\{c_\alpha\}$ coefficients fulfill convex conditions $\{c_\alpha > 0; \forall\, \alpha\} \wedge \{\sum_a c_\alpha = 1\}$. With this linear combination of υ matrices, the final similarity matrix $\mathbf{Z}$ can be interpreted as a possible representation of the molecular set studied and used in the multilinear regression model defined in eq 10. The optimal QSAR model is constructed choosing the set of $\{c_\alpha\}$ coefficients that best correlates with a given property of the molecular set $\mathbf{M}$. The $\{c_\alpha\}$ coefficients have been optimized in order to maximize the Q^2 statistical coefficient. This development, which can be referred to as tuned QSAR (TQSAR), adds a new tool for the application of MQSM to QSAR studies.

Before computing the set of $\{c_\alpha\}$ coefficients, all similarity matrices are normalized as follows

$$Z_{IJ}^{(N)} = Z_{IJ} \left(\sum_{KL} Z_{KL} \right)^{-1} \tag{14}$$

in order to obtain correctly scaled coefficients. Then $\{c_\alpha\}$ optimized coefficients will represent the contribution weights which every MQSM in the TQSAR model possesses.

Reduction of the Dimension. Before performing the multilinear regression analysis described in eq 10, a transformation of the similarity matrix can be carried out in order to discard redundant variables and to reduce the problem dimension. One of the most usual procedures employed in QSAR studies consists in transforming the variables matrix into a simplified matrix with the main factors or components. The method employed in this work is the classical scaling.[30] This is a well-established statistical tool which transforms the variables present in the data set into a series of orthogonal components. It can be shown that a p-dimensional exact solution always exist, where $p \leq m - 1$.[31] Generally, only three or four principal coordinates (PCs) are necessary to describe accurately the property studied. In this way, a new transformed set of point-molecules is obtained, $\mathbf{Y}(\Omega_\alpha) = \{\mathbf{y}_1, \mathbf{y}_2, \ldots \mathbf{y}_m\}$, which are p-dimensional vectors, p being the number of principal components.

The selection of the multilinear regression parameters is not a trivial task. An obvious selection consists of choosing the kth first eigenvectors arranged in descending order of the corresponding eigenvalues. This selection provides the best fitting k-dimensional subspace, where the sum of the interpoint distances is maximal. But there is not a unique variables selection method available. Here it has been used the *most predictive variables method* recently introduced by Cuadras.[32,33] This method selects as variable predictor, $k \leq p$ columns of the configuration matrix arranged in descending order of absolute correlation with the data, i.e.,

$$\chi^2(\pi, \mathbf{v}_{I_1}) > \ldots > \chi^2(\pi, \mathbf{v}_{I_k}) \tag{15}$$

where $\mathbf{v}_I$ is the Ith principal coordinate, $\pi = \{\pi_I;\ \forall\, I \in \mathbf{M}\}$ and χ^2 are defined as

$$\chi^2(\pi, \mathbf{v}_I) = \frac{(\pi^{\mathrm{T}} \mathbf{v}_I)^2}{\sum_J (\pi_J - \bar{\pi})^2 \lambda_J} \tag{16}$$

λ_J being the eigenvalue of the J axis. A coefficient of determination can be defined as

$$X_{(k)}^2 = \sum_{\alpha=1}^{k} \chi^2(\pi, \mathbf{v}_{J_\alpha}) \tag{17}$$

A maximal $X_{(k)}^2$ value has been used as a way to choose the k predictor variables. To avoid finding relationships including variables with high correlations and very low eigenvalues (a sign of noise parametrization), a *parameter filter* has been implemented in order to reject all those eigenvectors with eigenvalues less than 1% of the total variance.

Computational Scheme. Figure 1 describes the most relevant steps followed when a TQSAR study is performed over a molecular training set using MQSM. The first step consists of constructing the *promolecular* ASA density function for each molecule. To attain this goal, it is necessary to know two data sets: the molecular geometries and the atomic basis set functions. Molecular geometries can be obtained experimentally or as a result of computational procedures. On the other hand, the atomic data sets of coefficients and exponents required to generate *promolecular* ASA density functions were already computed in a previous work[12] and are available in ref 34. In the present study, *promolecular* ASA densities have been constructed using the following rule: one function at H, three functions for C, N, and O, and four functions for Cl.

The following step corresponds to MQSM computation and the search for the optimal molecular superposition. This

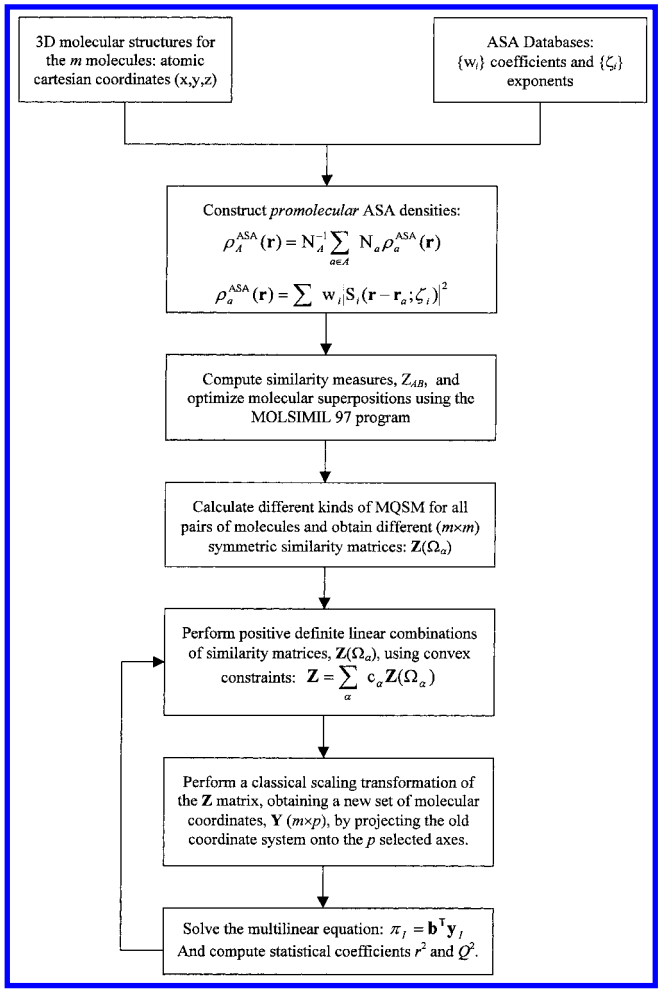

Figure 1. Computational flowchart describing the most relevant steps in a TQSAR study using MQSM.

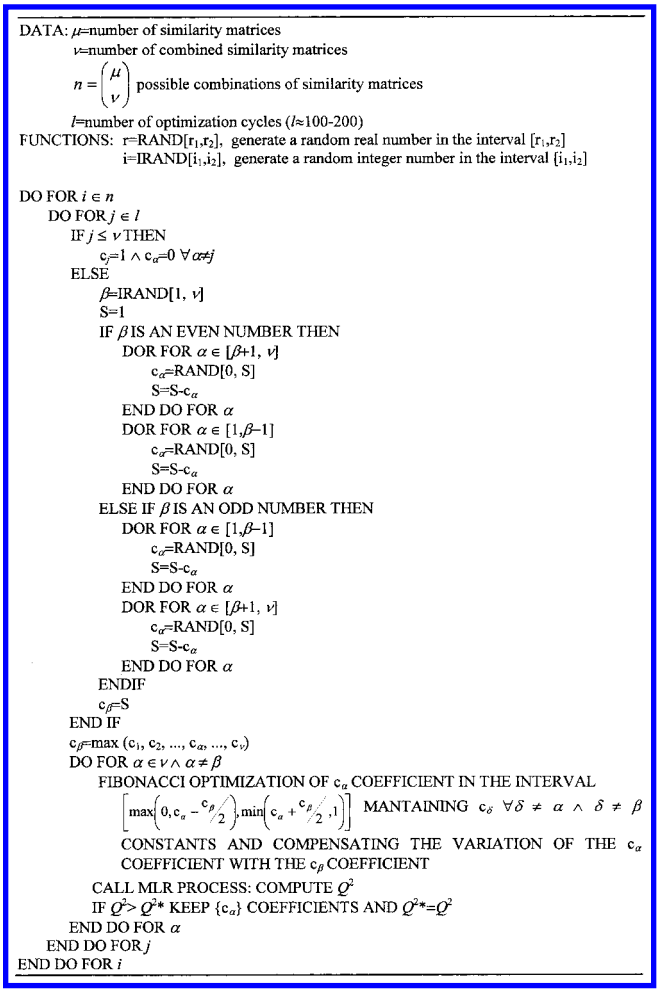

Figure 2. Computational algorithm describing the optimization process followed in the search for the linear combination of similarity matrices which produces the best TQSAR model.

process is carried out using the MOLSIMIL 97 program,[35] which includes the molecular superposition algorithm described above. The search for the molecular alignment, corresponding to the maximum similarity value for each MQSM, is one of the most time-consuming parts of the similarity studies.

After computing different kinds of MQSM for all the pairs of molecules, μ similarity matrices, $\mathbf{Z}(\Omega_\alpha)$, are obtained, and the process of constructing a TQSAR model can begin. This step is performed using the TQSAR-SIM program,[36] which is an evolution of previous multilinear regression programs developed in our laboratory. This new program includes the computational algorithm described in Figure 2 and used to optimize the set of $\{c_\alpha\}$ coefficients which define the TQSAR model. The objective is to maximize the Q^2 statistical coefficient by means of a combination of Monte Carlo[37] and Fibonacci[38] techniques.

The optimization is performed in the following way. First, a random sequence of $\{c_\alpha\}$ coefficients is generated using a Monte Carlo method. Then, starting from this sequence, a one-dimensional Fibonacci search is carried out for each c_α coefficient separately. This basic cycle is repeated until a satisfactory TQSAR model is produced with a maximum value of the statistical parameters. In addition, the process is repeated for all possible combinations of the $\mathbf{Z}(\Omega_\alpha)$ matrices in order to find the best linear combination. These combinations are generated using a nested summation symbol (NSS)[39,40] parallelizable algorithm.

Two parameters have to be fixed before running the TQSAR-SIM program: the number of matrices involved in the linear combination, ν, and the number of optimization cycles, l. It should be noted that the optimization process has $\nu-1$ degrees of freedom due to the fact that one coefficient, denoted in Figure 2 as c_β, depends on the remaining coefficients to satisfy the convex constraint

$$c_\beta = 1 - \sum_{\alpha \neq \beta} c_\alpha \qquad (18)$$

In the current and preliminary version of the TQSAR-SIM program, a great deal of computer time was needed to execute this optimization algorithm. However, some present studies in our laboratory are oriented to determine *a priori* which MQSM are the most important for a given molecular set and, in this way, reduce the number of combinations of similarity matrices.[41] On the other hand, the performance of the optimization search could be improved by further useful modifications: using the elementary Jacobi rotations technique to find the optimal sequence of $\{c_\alpha\}$ coefficients, for instance.

Following the flowchart shown in Figure 1, the next step is to reduce the variable dimension of the $\mathbf{Z}$ matrix by means of a classical scaling analysis, which transforms the molecular points into a set of orthogonal components. Then, solving the multilinear equation defined in the expression 10 using

Table 1. Structures and Topoisomerase-I Inhibition Activity for CPT Analogues

substituents	compound (identification no.)	log ($1/IC_{50}$)
10,11-OCH_2O-, 20(*S*)	10,11-methylenedioxy-20(*S*)-CPT (**28**)	1.569
9-CH_3, 20(*S*)	9-methyl-20(*S*)-CPT (**18**)	1.420
9-NH_2-10,11-OCH_2O-, 20(*S*)	9-amino-10,11-methylenedioxy-20(*S*)-CPT (**6**)	1.319
9-Cl-10,11-OCH_2O-, 20(*S*)	9-chloro-10,11-methylenedioxy-20(*S*)-CPT (**30**)	1.215
9-Cl, 20(*S*)	9-chloro-20(*S*)-CPT (**16**)	1.066
10-OH, 20(*S*)	10-hydroxy-20(*S*)-CPT (**3**)	0.975
9-NH_2, 20(*S*)	9-amino-20(*S*)-CPT (**9**)	0.955
10-NH_2, 20(*S*)	10-amino-20(*S*)-CPT (**10**)	0.854
10-Cl, 20(*S*)	10-chloro-20(*S*)-CPT (**17**)	0.851
10-NO_2, 20(*S*)	10-nitro-20(*S*)-CPT (**13**)	0.197
20(*S*)	20(*S*)-CPT (**1**)	0.169
9-OH, 20(*S*)	9-hydroxy-20(*S*)-CPT (**15**)	0.059

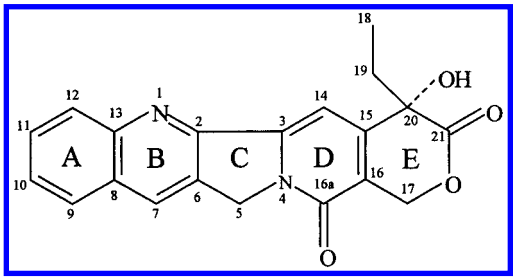

Figure 3. Structure of the CPT molecule.

the most predictive dimensions method, a QSAR model is constructed and validated for the r^2 and the Q^2 statistical parameters.

RESULTS: A PRACTICAL EXAMPLE

The theoretical methods previously described for the computation of MQSM have been employed in a TQSAR study of a family of antitumor agents: the CPT molecule and analogues. CPT is a natural product isolated from the fruit of *Camptotheca acuminata*, a tree indigenous to China, which belongs to the *Nyssaceae* family.[18] The structure of the CPT molecule is shown in Figure 3, and analogues are constructed from this structure using the substituents described in Table 1. These active compounds are cytotoxic drugs with T-I inhibition activity and have an asymmetric atom, denoted by the 20(*S*) atomic configuration. In addition, they can be considered as rigid molecules without important conformational problems, and consequently the MQSM study can be performed taking computationally frozen structures. In order to perform calculations in this paper, the minimal energy conformation has been taken into account for all structures. Another reason why this molecular set has been chosen can be found in the impossibility of establishing *a priori* relationships between antitumor activity and the substituents of the chemical structures.

Preliminary Structural Study. The generation of 3D molecular structures is one of the most important problems related to MQSM studies. The best structures, if they are available, are the experimental ones. To this effect, a previous search for experimental structures for CPT analogues in the Cambridge Structural Database (CSD)[42] was performed, and the crystallographic structure of the 5beta-hydroxymethyl camptothecin molecule (5(*S*)-CH_2OH-20(*S*)-CPT)[43] was found. This X-ray crystallographic structure, denoted as YIZYEL in the CSD reference code, has been employed as a guide for a theoretical study to determine which method can be used to optimize the structures of CPT analogues presented in Table 1. In fact, this structural study shows another possible application example of MQSM, in a similar manner to previous works where a comparison of different calculation levels[44] or basis set[45] was performed. Three different methodologies have been compared: molecular mechanics, semiempirical, and *ab initio*, using the programs HyperChem,[46] AMPAC 5.0,[47] and Gaussian 94,[48] respectively. Four different geometrical energetic minimizations of the YIZYEL molecule have been carried out, starting in all cases from the X-ray crystallographic structure. The first corresponds to a molecular mechanics calculation using the MM+ force field, which is an extension of the MM2 technique developed by Allinger and co-workers.[49] AM1[50] and PM3[51] semiempirical hamiltonians have been used to optimize the 3D molecular structure of the YIZYEL molecule. Finally, an *ab initio* calculation at the HF/3-21G[52] level of theory has been performed. In order to compare the four optimized geometries of the YIZYEL molecule with respect to the X-ray crystallographic one in a general way, overlap-like MQSM are computed using the MOLSIMIL 97 program and *promolecular* ASA densities described above. Carbó indices for all pairs of molecular structures are presented in Table 2. The highest Carbó index obtained is 0.783 and corresponds to the pair X-ray − HF/3-21G. This result indicates that these two geometries are the most similar of the set. On the other hand, the lowest Carbó index value is 0.395 and corresponds to the molecular superposition X-ray − MM+. The molecular mechanics calculation is the fastest approach, but the use of empirical energy functions is insufficient to obtain molecular geometries close to the experimental ones. In order to illustrate the difference between both structures, Figure 4 depicts the optimized molecular superposition of the pair X-ray − MM+. By inspection of this figure, it should be noted that the most important difference between both geometries lies in rings A and B, according to Figure 3.

Table 2. Carbó Index Values for the YIZYEL Molecule Used To Compare Different Optimization Methodologies

	MM+	AM1	PM3	3-21G	X-ray
MM+	1	0.452	0.436	0.487	0.395
AM1	0.452	1	0.652	0.606	0.622
PM3	0.436	0.652	1	0.564	0.589
3-21G	0.487	0.606	0.564	1	0.783
X-ray	0.395	0.622	0.589	0.783	1

This theoretical study demonstrates that the best geometry is the *ab initio* one, but this calculation is not viable when

Table 3. Square Correlation Coefficients and Square Predictive Correlation Coefficients for All the Different Types of MQSM

similarity matrices	abbreviation	3 PCs Q^2	3 PCs r^2	4 PCs[a] r^2
overlap	OVE	0.003	0.467	0.620
Coulomb	COU	0.078	0.432	0.457
gravitational	GRV	<0	0.599	0.672
ρ_C=10,11-OCH_2O-, 20(*S*)-CPT	T**28**	<0	0.476	0.477
ρ_C=9-CH_3, 20(*S*)-CPT	T**18**	<0	0.525	0.638
ρ_C=9-NH_2-10,11-OCH_2O-, 20(*S*)-CPT	T**06**	0.176	0.481	0.537
ρ_C=9-Cl-10,11-OCH_2O-, 20(*S*)-CPT	T**30**	<0	0.293	0.361
ρ_C=9-Cl, 20(*S*)-CPT	T**16**	<0	0.289	0.308
ρ_C=10-OH, 20(*S*)-CPT	T**03**	<0	0.504	0.504
ρ_C=9-NH_2, 20(*S*)-CPT	T**09**	0.023	0.436	0.498
ρ_C=10-NH_2, 20(*S*)-CPT	T**10**	<0	0.447	0.505
ρ_C=10-Cl, 20(*S*)-CPT	T**17**	<0	0.463	0.479
ρ_C=10-NO_2, 20(*S*)-CPT	T**13**	<0	0.476	0.506
ρ_C=20(*S*)-CPT	T**01**	<0	0.487	0.498
ρ_C=9-OH, 20(*S*)-CPT	T**15**	0.361	0.558	0.643

[a] Q^2 for four PCs calculations was found <0 in all cases.

Table 4. Square Correlation Coefficients and Square Predictive Correlation Coefficients for Different TQSAR Models

υ[a]	k[b]	optimal PCs	TQSAR model	Q^2	r^2
2	3	1, 6, 8	0.376×T**18** + 0.624×T**06**	0.730	0.843
	4	1, 6, 8, 5	0.378×T**18** + 0.622×T**06**	0.755	0.869
3	3	1, 8, 7	0.389×T**18** + 0.521×T**06** + 0.090×T**03**	0.842	0.884
	4	1, 8, 7, 5	0.355×T**18** + 0.558×T**06** + 0.087×T**03**	0.858	0.899
4	3	1, 8, 9	0.426×COU + 0.285×T**18** + 0.269×T**06** + 0.020×T**01**	0.866	0.928
	4	1, 8, 9, 3	0.468×COU + 0.264×T**18** + 0.248×T**06** + 0.020×T**01**	0.842	0.916

[a] Number of similarity matrices. [b] Number of PCs.

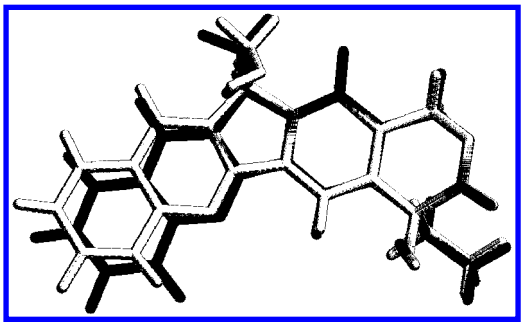

Figure 4. Molecular superposition of the pair X-ray − MM+. X-ray in light color and MM+ in dark color.

a large number of large-sized molecules is studied due to its high computational cost. This effect is corroborated by evaluating the primitive Gaussian functions employed in this *ab initio* computation: 288 basis functions and 474 primitive GTO. The second closest geometry to the X-ray is the semiempirical AM1 structure as shown in Table 2. Due to these facts, 3D molecular structures for CPT analogues have been generated with the AMPAC 5.0 program, using the AM1 methodology, and starting from the X-ray crystallographic configuration of the YIZYEL molecule.

QSAR Results Using one Similarity Matrix. Fifteen different similarity matrices have been computed for the set of 12 CPT analogues: overlap-like (OVE), Coulomb-like (COU), gravitational-like (GRV), and 12 TMQSM. These TMQSM are computed by substituting the positive definite operator Ω_α by the density function of each molecule in the studied set. All these similarity matrices can be downloaded from the WWW site specified in ref 53. For all these similarity matrices, a classical scaling analysis has been carried out keeping the first three and four components to solve the multilinear equation and construct a QSAR model. Square correlation coefficients and square predictive correlation coefficients for all of them are listed in Table 3.

The Q^2 negative values which appear in Table 3 indicate a poor predictive power for the QSAR model generated. It should be noted that these negative values are possible due to the Q^2 definition as expressed in the eq 11. The best result corresponds to the TMQSM computation using three PCs and the CPT analog T**15** as the operator Ω_α, obtaining a Q^2 of 0.361 and a r^2 of 0.558. Other operators present positive prediction coefficients, such as OVE, COU, T**06**, and T**09**. The rest bear Q^2 negative values. When four PCs are used, all the correlation coefficients r^2 are improved, but on the other hand overall negative prediction coefficients are obtained.

Tuned QSAR Results. The computational scheme described in Figure 1 has been followed to obtain different TQSAR models using the 15 similarity matrices calculated in the previous section ($\mu = 15$). Statistical results obtained for the TQSAR models combining two ($\upsilon = 2$), three ($\upsilon = 3$), and four ($\upsilon = 4$) matrices are listed in Table 4.

In contrast to the low values of the square predictive correlation coefficients obtained using only one similarity matrix, the TQSAR models produce more accurate results. This study provides evidence for the great potential of the combination of different MQSM for the development of TQSAR models. The TMQSM operators T**18** and T**06** are the basis of the constructed TQSAR models, possessing the major weight, although for the models obtained combining four matrices, the weight is shifted to the COU similarity matrix, and the T**01** TMQSM operator emerges with an almost negligible weight. This fact corrects the apparent outlier nature of compound **1** (the nonsubstituted, original camptothecin molecule) appearing in the models with three

Table 5. Topoisomerase-I Inhibition Activity for CPT Analogues: (a) Experimental, (b) Predicted Using Three Similarity Matrices and Four PCs, and (c) Predicted Using Four Similarity Matrices and Three PCs

compound	y^a	y^b	y^c
28	1.569	1.637	1.391
18	1.420	1.396	1.449
6	1.319	1.330	1.212
30	1.215	1.131	1.440
16	1.066	1.024	1.287
3	0.975	0.958	0.720
9	0.955	0.997	0.956
10	0.854	0.611	0.648
17	0.851	0.642	1.041
13	0.197	0.249	0.136
1	0.169	0.676	0.452
15	0.059	−0.061	0.080

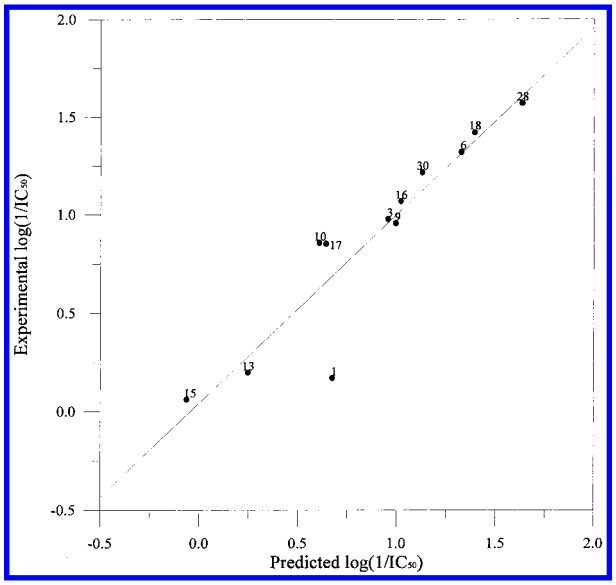

Figure 5. Representation of predicted vs experimental values of the T-I inhibition activity for the set of 12 CPT analogues using three similarity matrices and four PCs.

matrices, as will be explained below.

A good regression model is obtained when three matrices (T**18**, T**06**, and T**03**) and four PCs are employed, yielding $Q^2 = 0.858$ and $r^2 = 0.899$ values. The representation of the experimental activities vs the predicted ones (listed in Table 5) for this model is shown in Figure 5, where a good fitting of the activities for all the compounds is obtained, except for compound **1**. When four matrices are combined, the results are not significantly improved in relation to the amount of CPU time required, but the full set of compounds (included the CPT molecule) are well-correlated. The predicted T-1 values obtained using four matrices and three PCs are included in Table 5 and represented in Figure 6. In this analysis, predictive regression coefficient (Q^2) is 0.866. In fact, these results confirm that TQSAR models have better predictive capabilities than a simple QSAR model.

CONCLUSIONS

MQSM theory and methodology applied to 3D QSAR studies is well established today. TQSAR analysis of CPT analogues using the T-I inhibition activity described herein demonstrates that MQSM can be a useful tool to predict the activities of therapeutic agents. In addition, the present

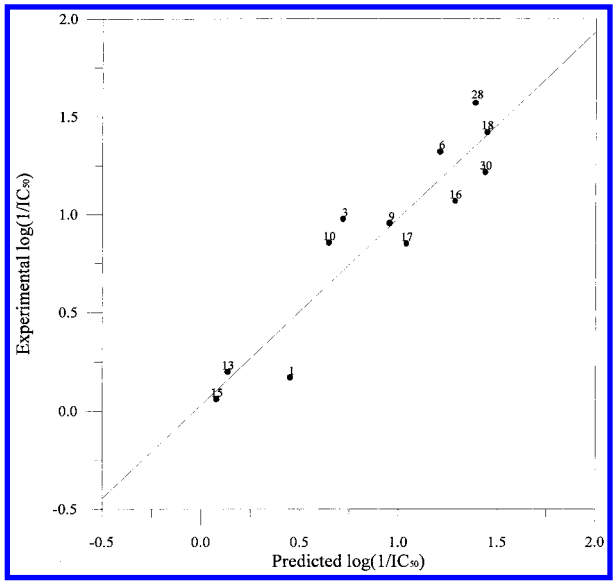

Figure 6. Representation of predicted vs experimental values of the T-I inhibition activity for the set of 12 CPT analogues using four similarity matrices and three PCs.

results suggest that linear combinations of similarity matrices using convex set constraints provide an effective means of generating accurate QSAR models. It must be emphasized that by using a convex set formalism in MQSM studies, more consistent QSAR models can be generated. The results obtained open the path for future TQSAR analysis using linear combinations of different kinds of MQSM to construct a QSAR model based on optimal discrete molecular descriptors of quantum mechanical origin.

ACKNOWLEDGMENT

This research was partially supported by a CICYT Grant SAF 96-0158 and the *Fundació Maria Francisca de Roviralta*. One of us (L.A.) benefits from a fellowship at the Spanish *Ministerio de Educación y Cultura*. Thanks are also due to the *Centre de Supercomputació de Catalunya (CESCA)* and the *Centre Europeu de Paral·lelisme de Barcelona (CEPBA)* for a generous amount of computation time. The authors thank the referees for their constructive criticism, which has improved in many aspects this work.

REFERENCES AND NOTES

(1) Kubinyi, H., Ed. *3D QSAR in Drug Design: Theory Methods and Applications*; ESCOM Science Publishers B.V.: Leiden, The Netherlands, 1993.

(2) *Molecular Similarity in Drug Design*; Dean, P. M., Ed.; Blackie Academic & Professional: London, 1995.

(3) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.

(4) Hermann, R. B.; Herron, D. K. OVID and SUPER: two overlap programs for drug design. *J. Comput.-Aided Mol. Design* **1991**, *5*, 511−524.

(5) Good, A. C.; So, S. S.; Richards, W. G. Structure-Activity Relationships from Molecular Similarity Matrices. *J. Med. Chem.* **1993**, *36*, 433−438.

(6) Mestres, J.; Rohrer, D. C.; Maggiora, G. M. A molecular field-based similarity approach to pharmacophoric pattern recognition. *J. Mol. Graphics* **1997**, *15*, 114−121.

(7) Carbó, R.; Besalú, E.; Amat, L.; Fradera, X. Quantum molecular similarity measures (QMSM) as a natural way leading towards a theoretical foundation of quantitative structure-properties relationships (QSPR). *J. Math. Chem.* **1995**, *18*, 237−246.

(8) Fradera, X.; Amat, L.; Besalú, E.; Carbó-Dorca, R. Application of Molecular Quantum Similarity to QSAR. *Quant. Struct.-Act. Relat.* **1997**, *16*, 25−32.
(9) Lobato, M.; Amat, L.; Besalú, E.; Carbó-Dorca, R. Structure-Activity Relationships of a Steroid Family using Quantum Similarity Measures and Topological Quantum Similarity Indices. *Quant. Struct.-Act. Relat.* **1997**, *16*, 465−472.
(10) Constans, P.; Carbó, R. Atomic Shell Approximation: Electron Density Fitting Algorithm Restricting Coefficients to Positive Values. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1046−1053.
(11) Constans, P.; Amat, L.; Fradera, X.; Carbó-Dorca, R. Quantum Molecular Similarity Measures (QMSM) and the Atomic Shell Approximation (ASA). In *Advances in Molecular Similarity*; Carbó-Dorca, R., Mezey, P. G., Eds.; JAI Press Inc.: Greenwich, CT, 1996; Vol. 1, pp 187−211.
(12) Amat, L.; Carbó-Dorca, R. Quantum Similarity Measures under Atomic Shell Approximation: First Order Density Fitting Using Elementary Jacobi Rotations. *J. Comput. Chem.* **1997**, *18*, 2023−2039.
(13) Constans, P.; Amat, L.; Carbó-Dorca, R. Toward a Global Maximization of the Molecular Similarity Function: Superposition of Two Molecules. *J. Comput. Chem.* **1997**, *18*, 826−846.
(14) Amat, L.; Carbó, R.; Constans, P. Algorisme d'optimització global de les mesures de semblança quàntica molecular. *Sci. Gerund.* **1996**, *22*, 109−121.
(15) Carbó-Dorca, R. Tagged Sets, Convex Sets and Quantum Similarity Measures. *J. Math. Chem.* **1997**, *22*, 143−148.
(16) Carbó-Dorca, R. Fuzzy Sets and Boolean Tagged Sets; Vector Semispaces and Convex Sets; Quantum Similarity Measures and ASA Density Functions; Diagonal Vector Spaces and Quantum Chemistry. Technical report: IT-IQC-20-97.
(17) Hsiang, Y. H.; Hertzberg, R.; Hecht, S.; Liu, L. F. Camptothecin Induces Protein-linked DNA Breaks via Mammalian DNA Topoisomerase I. *J. Biol. Chem.* **1985**, *260*, 14873−14878.
(18) Wall, M. E.; Wani, M. C. Camptothecin and Analogues. In *Cancer Chemotherapeutic Agents*; Foye, W. O., Ed.; ACS Professional Reference Book: Washington, 1995; Chapter 7, pp 293−310.
(19) *Concepts and Applications of Molecular Similarity*; Johnson, M. A., Maggiora, G., Eds.; John Wiley & Sons, Inc.: New York, 1990.
(20) *Molecular Similarity and Reactivity: From Quantum Chemical to Phenomenological Approaches*; Kluwer Academic: Amsterdam, 1995.
(21) *Advances in Molecular Similarity*; Carbó-Dorca, R., Mezey, P. G., Eds.; JAI Press Inc.: Greenwich, CT, 1996; Vol. 1.
(22) *Advances in Molecular Similarity*; Carbó-Dorca, R., Mezey, P. G., Eds.; Vol. 2. In press.
(23) Carbó, R.; Besalú, E.; Amat, L.; Fradera, X. On quantum molecular similarity measures (QMSM) and indices (QMSI). *J. Math. Chem.* **1996**, *19*, 47−56.
(24) Carbó, R.; Leyda, L.; Arnau, M. How Similar is a Molecule to Another? An Electron Density Measure of Similarity between Two Molecular Structures. *Int. J. Quantum Chem.* **1980**, *17*, 1185−1189.
(25) Mestres, J.; Solà, M.; Duran, M.; Carbó, R. On the Calculation of *Ab Initio* Quantum Molecular Similarities for Large Systems: Fitting the Electron Density. *J. Comput. Chem.* **1994**, *15*, 1113−1120.
(26) Cioslowski, J.; Piskorz, P.; Rez, P. Accurate analytical representations of the coreelectron densities of the elements 3 through 118. *J. Chem. Phys.* **1997**, *106*, 3607−3612.
(27) von Neumann, J. *Mathematical Foundations of Quantum Mechanics*; Princeton Univesrity Press: Princeton, NJ, 1955.
(28) Carbó, R.; Calabuig, B. Quantum Similarity Measures, Molecular Cloud Descriptors, and Structure-Properties Relationships. *J. Chem. Inf. Comput. Sci.* **1992**, **32**, 600−606.
(29) Montgomery, D. C.; Peck, E. A. *Introduction to Linear Regression Analysis*; John Wiley & Sons, Inc.: New York, 1992.
(30) Cox, T. F.; Cox, M. A. A. *Multidimensional Scaling*; Chapman & Hall: London, 1994.
(31) Mardia, K. V.; Kent, J. T.; Bibby, J. M. *Multivariate Analysis*; Academic Press: London, 1979.
(32) Cuadras, C. M.; Arenas, C. A distance based regression model for prediction with mixed data. *Commun. Statist.-Theory Meth.* **1990**, *19*(6), 2261−2279.
(33) Cuadras, C. M.; Arenas, C.; Fortiana, J. Some computational aspects of a distance-based model for prediction. *Commun. Statist.-Simula.* **1996**, *25*(3), 593−609.
(34) ASA coefficients and exponents can be seen and downloaded from?the WWW site: http://iqc.udg.es/cat/similarity/ASA/funcset.html.
(35) Amat, L.; Constans, P.; Besalú, E.; Carbó-Dorca, R. MOLSIMIL 97; Institute of Computational Chemistry, University of Girona: Spain, 1997.
(36) Amat, L.; Robert, D.; Besalú, E. TQSAR-SIM; Institute of Computational Chemistry, University of Girona: Spain, 1997.
(37) Demidovich, B. P.; Maron, I. A. *Computational Mathematics*; Mir Publishers: Moscow, 1981.
(38) Pierre, D.A. *Optimization Theory with Applications*; John Wiley & Sons, Inc.: New York, 1969.
(39) Carbó, R.; Besalú, E. Definition, mathematical examples and quantum chemical applications of nested summation symbols and logical Kronecker deltas. *Computers Chem.* **1994**, *18*, 117−126.
(40) Carbó, R.; Besalú, E. Definition and quantum chemical applications of nested summation symbols and logical functions: Pedagogical artificial intelligence devices for formulae writing, sequential programming and automatic parallel implementation. *J. Math. Chem.* **1995**, *18*, 37−72.
(41) Robert, D.; Carbó-Dorca, R. Analyzing the Triple Density Quantum Similarity Measures with the INDSCAL Model. *J. Chem. Inf. Comput. Sci.* Technical report: IT-IQC-24-**97**. In press.
(42) Allen, F. H.; Bellard, S.; Brice, M. D.; Cartwright, B. A.; Doubleday, A.; Higgs, H.; Hummelink, T.; Hummelink-Peters, B. G.; Kennard, O.; Motherwell, W. D. S.; Rodgers, J. R.; Watson, D. G. The Cambridge Crystal Data Centre: Computer-Based Search, Retrieval, Analysis and Display of Information. *Acta Crystallogr.* **1979**, *B35*, 2331−2339.
(43) Wang, H. K.; Liu, S. Y.; Hwang, K. M.; McPhail, A. T.; Lee, K. H. *Bioorg. Med. Chem. Lett.* **1995**, *5*, 77.
(44) Solà, M.; Mestres, J.; Carbó, R.; Duran, M. A comparative analysis by means of quantum molecular similarity measures of density distributions derived from conventional *ab initio* and density functional methods. *J. Chem. Phys.* **1996**, *104*, 636−647.
(45) Solà, M.; Mestres, J.; Oliva, J. M.; Duran, M.; Carbó, R. The Use of Ab Initio Quantum Molecular Self-Similarity Measures to Analyze Electronic Charge Density Distributions. *Int. J. Quant. Chem.* **1996**, *58*, 361−372.
(46) HyperChem, Release 3 for Windows. Molecular Modeling System. 1993 Hypercube, Inc. and Autodesk, Inc.
(47) AMPAC 5.0, 1994 Semichem, 7128 Summit, Shawnee, KS 66216.D.A.
(48) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Gill, P. M. W.; Johnson, B. G.; Robb, M. A.; Cheeseman, J. R.; Keith, T. A.; Petersson, G. A.; Montgomery, J. A.; Raghavachari, K.; Al-Laham, M. A.; Zakrzewski, V. G.; Ortiz, J. V.; Foresman, J. B.; Cioslowski, J.; Stefanov, B. B.; Nanayakkara, A.; Challacombe, M.; Peng, C. Y.; Ayala, P. Y.; Chen, W.; Wong, M. W.; Andres, J. L.; Replogle, E. S.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Binkley, H. S.; Defrees, D. J.; Baker, H.; Stewart, J. J. P.; Head-Gordon, M.; Gonzalez, C.; Pople, J. A. GAUSSIAN 94, Revision A.1; Gaussian, Inc.: Pittsburgh, PA, 1995.
(49) Allinger, N. L. Conformational analysis 130. MM2. A Hydrocarbon Force Field Utilizing V1 and V2 Torsional Terms. *J. Am. Chem. Soc.* **1977**, *99*, 8127−8134.
(50) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.
(51) Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods. I. Method. *J. Comput. Chem.* **1989**, *10*, 209−220.
(52) Binkley, J. S.; Pople, J. A.; Hehre, W. J. Self-Consistent Molecular Orbital Methods. 21. Small Split-Valence Basis Sets for First-Row Elements. *J. Am. Chem. Soc.* **1980**, *102*, 939−947.
(53) Molecular structures and similarity matrices can be seen and downloaded from the WWW site: http://iqc.udg.es/cat/similarity/QSAR/cpt.html.

CI9800108